ALAGAPPA UNIVERSITY

(Accredited with 'A' Grade by NAAC)

KARAIKUDI-630 003, TAMILNADU

DIRECTORATE OF DISTANCE EDUCATION

(Recognized by Distance Education Council (DEC), New Delhi)

M.Sc. (Zoology)

35043

BIOPHYSICS, BIOSTATISTICS AND BIOINFORMATICS

SYLABBI-BOOK MAPPING TABLE

CURRICULUM AND INSTRUCTUION

Syllabi	Mapping in Book

BLOCK-I: BIOPHYSICS

UNIT I	Pages 1-13
Introduction	
Structure and properties of atoms and molecules	
Chemical bonds	
Types and properties	
Polymerization of organic molecules.	
UNITS II	Pages 14-25
Laws of thermodynamics	
Principle and application	
Bio-energetics	
Coupling of chemical reactions	
Redox potential	
NADP/NADPH and Free energy	
UNIT III	Pages 26-37
Natural Radiations	
Properties of light	
Absorption of light	
Energy states of atoms & Spin property of electrons	
Ground state and exited state of atoms & Bio molecules	

BLOCK III: MEASURES OF CENTRAL TENDENCY AND MEASURE OF DISPERSION

BLOCK IV: BIOINFORMATICS

CONTENTS

BLOCK-I: BIOPHYSICS

35043-BIOPHYSICS, BIOSTATISTICS AND BIOINFORMATICS

BLOCK-I: BIOPHYSICS

UNIT I

1.1. Introduction :

Biophysics is an interdisciplinary science which applies the principles of physics to the study of biology in order to increase the understanding of biological systems. For the human being, biophysics can be thought of a providing a description of his whole physical system from the particular view of physics. The chief concern in the development of biophysics is that those skills should be acquired by people who start with the right intellectual approach, both physical and biological. The molecules which provide the required frame and dynamism for the life processes are made up of smaller units called atoms. For example, a molecule of glucose is made up of 6 carbon, 6 oxygen and 12 hydrogen atoms. The most abundant atoms are those of carbon, hydrogen, oxygen and nitrogen, which together constitute about 99% of the total atoms in most living systems. Calcium, phosphorus, magnesium, potassium, sodium, sulphur, iron and chlorine atoms also form a significant proportion of the biological molecules. This unit is meant to study the structure and properties of an atom and molecules with their atomic strength.Then understanding the formation and orbitals of atoms and molecules with bond types and polymerization of organic molecules. To study the principle and application of laws of thermodynamics and the free energy from electromagnetic waves. Finally to know the natural radiations, absorption of light, energy states of atoms, spectroscopy and measurement of radio activity.

1.2 Structure and Properties of Atoms and Molecules:

After the existence of atomic particles was established, the study of the structure of these particles was undertaken by several physical scientists. Although, at one time the atoms were conceived to be the smallest particles, sub-atomic particles were recognized in due course of time. The three fundamental sub-atomic particles are proton, neutron and electron. In addition to these fundamental particles, about 35 or more other atomic particles are also known to exist. Many of them are, however, extremely unstable and they merely represent a bundle of energy.

Some of the stable particles other than fundamental particles are positron, photon, neutrino, graviton and antiproton. These particles are however, of little importance in the study of biochemistry because their existence is rarely encountered during the conventional study of the biological systems.

Regarding the arrangement of fundamental particles inside an atom, several models were proposed. However, the most satisfactory model was proposed by Earnest Rutherford in 1911. This is accepted even today with some modifications. Accordingly an atom is made up of a central nucleus containing positively charged protons and neutral neutrons and one or more orbits or shells of negatively charged electrons, moving around the nucleus (Fig. 1). For many years, it was customary to describe Rutherford's model as a planetary model; the model being similar to the sun's planetary system, in which the planets move in orbit around the sun. However, the elements in a planetary system are not charged, while the sub-atomic particles protons and electrons bear positive and negative charges respectively.

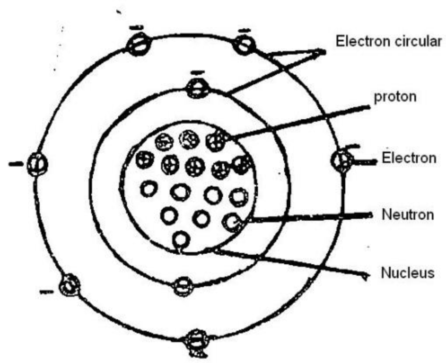


Fig. 1 : Structure of an atom

Hydrogen atom is the simplest atom which consists of one proton, one neutron and one electron. In this atom, the one electron is situated in one orbit or shell around the nucleus. In helium also two electrons are situated in the single shell. However, in other elements the electrons are arranged in several shells. Thus a .neon atom has two shells of 2 and 8 electrons (total 10) and argon has three shells of 2, 8 and 8 (total 18) electrons.

These orbits or shells may never have more electrons than a certain maximum. This maximum value depends upon the position of the shell and is represented by the term $2N^2$, where N is the order (first, second, third etc.) of the shell. Thus the maximum number of electrons in fourth shell of an atom would be $2 \times 4^2 = 32$. When this maximum number of electrons for a particular shell is reached, the shell is said to be saturated. The elements whose outer most shells are saturated with electrons, are relatively inert and do not participate in the chemical reactions under normal conditions. The elements in this list (helium, neon, argon, krypton, xenon etc.,) are gases at normal temperature and they are called noble gases, because of their inertness. The elements whose atoms have electrons one more or one less (or even higher values) than the inert gas configuration in their outer most shell, are chemically active. This chemical activity may be interpreted as a tendency of those atoms to acquire noble configuration by accepting or losing one electrons.

The electron shells 1, 2, 3, 4 etc., are also called K.L.M.N. etc., shells. Each shell is further divided into sub-shells (Table .1). The maximum number of electrons in the first shell is 2

and it has only one sub-shell called sub-shell. The second shell has two sub-shells, s and p and third shell has three sub-shells s, p and d and so on.

Table-1 : **Electrons distribution in shell and sub-shells**

Shell No. (n)	1	2		3				4 etc.,	
Shell name	K	L		M				N	
Sub-shells	0	0	1	0	1	2	0	2	3
No. of electrons	s	s	p	s	p	d	p	d	f
In sub-shells	2	2	6	2	6	10	6	10	14
Total no. of Electron in a shell	2	8		18				32	

The mass of an atom depends entirely upon its nucleus. A neutron has nearly the same mass as a proton. In absolute terms each proton or neutron weighs 1.66043 X 10~249. The mass of an electron is negligible, about 1/1823 of the mass of a proton or neutron. The mass of a proton or neutron is called the atomic mass unit (amu). This, if a carbon atom contains 6 neutrons, 6 protons and 6 electrons, its amu will be equivalent to the neutron + proton i.e. 6+6 = 12. This amu is also the atomic weight of the element of protons and neutron each taken as a unit weight. The number of protons on an atom is called the atomic number of the atom. Thus, the atom is fixed; it may vary sometimes giving rise to different species of the same atom. The different atomic species, having the same atomic number (as they have same proton number), but different atomic weight (as they have different neutron numbers) are called isotopes. Thus a hydrogen atom which contains only one neutron. The number of protons and electrons in both species of the hydrogen is same.

In an atom the number of electrons equals the number of protons, $N_e = N_p$. It is possible to have a system with N_e not equal to N_p, but then it is not called an atom; it is called an ion. On earth most of the matter we come in contact with is made of atoms. However, in the sun and the stars most of the matter is in the form of ions. It is probable that most of the matter in the universe is in the form of ions. The number of neutrons in the nucleus is in many cases approximately equal to the number of protons. A small number of atoms may combine to form a molecule such as water (H_2O) or carbon dioxide (CO_2. The arrangements and rearrangements of atoms in molecules form the domain of chemistry. A solid consists of a regular array or lattice containing a very large number molecules or atoms. A typical number of atoms in a piece of

matter on a human scale is 10^{24}. Physical properties of matter include hardness, malleability, color, and melting point. The important point here is that chemical and physical properties of matter depend only on the electron cloud surrounding the atoms. Therefore they depend only on the number of protons in the nucleus. This number is symbolized by Z, and is called the "atomic number"

Size : The atom is about 10^{-10} meters (or 10^{-8} centimeters) in size. This means a row of 10^{8} (or 100,000,000) atoms would stretch a centimeter, about the size of your fingernail. Atoms of different elements are different sizes, but 10^{-10} m can be thought of as a rough value for any atom. It is also a good approximation to think of atoms as spherical in shape, although they are not always so. The atom with the smallest mass is the hydrogen atom; its mass is about 10^{-27} kg. The masses of other atoms go up to about 200 times this. The nucleus of an atom is about 10^{-15} m in size; this means it is about 10^{-5} (or 1/100,000) of the size of the whole atom. A good comparison of the nucleus to the atom is like a pea in the middle of a racetrack. (10^{-15} m is typical for the smaller nuclei; larger ones go up to about 10 times that.)

Mass : Although it is very small, the nucleus is massive compared to the rest of the atom. Typically the nucleus contains more than 99.9% of the mass of the atom. (Hence the numbers given above for masses of atoms also apply approximately to the nucleus alone.) Nuclei are usually spherical in shape, although some are spheroidal (egg-shaped).

Subatomic particles**:** The nucleus is made up of protons and neutrons bound together by attractive forces. The outer volume of the atom (which means most of the atom) is occupied by electrons. An electron itself is small (its size is not known, but we do know that it is smaller than a nucleus), but it occupies the space of the atom by constantly whirling around in a kind of orbit around the nucleus.The proton and neutron are spherical, about 10^{-15} m in radius. The proton and neutron have almost the same mass - the neutron's is slightly larger. These masses are more than 2000 times the mass of the electron. That is why the nucleus has most the atom's mass

Forces inside the atom**:** The electron has a negative electric charge; the proton has a positive electric charge of exactly the same strength; the neutron has no electric charge. Like charges repel and unlike charges attract. The nucleus is a dense ball of positive charge in the center of the atom and it exerts an attractive force on the electrons, thereby holding them as part of the atom. In other words, atoms would not exist if it were not for the electric force.

Stability of the atom : Why don't the electrons fall into the nucleus under the influence of this force? The atom looks superficially like the solar system, with its planets in orbit around the central sun. But the reason for the atom's stability -- the fact that the electron's orbit does not collapse -- lies in the fundamental nature of quantum mechanics, the science that supersedes Newton's mechanics in the world of the atom. The electron can exist only in one of a discrete set of "energy states", and the lowest energy state is stable. The electron jumps from one state to another when it receives or emits a quantum of energy in the form of light (or other form of electromagnetic radiation)

The nuclear force: Inside the nucleus all the electric forces are repulsive because the protons repel each other, and the neutrons don't feel any electric force. How then is the nucleus held together? There is another force, called the nuclear force that is mostly attractive and acts between two protons, between two neutrons, and between a neutron and proton. In nuclei this force is stronger than the repulsive electric force and so nuclei are held together.

I.3. Chemical bonds – types and properties:

Several atoms are assembled and held together to form thousands of molecules which take part in the building and function of physical and biological systems. The formation of bond between two atoms is due to some redistribution or regrouping of electrons to form a more stable configuration. The important types of chemical bonds are described below:

1.3.1. Ionic bonds:

Ionic bonds are the properties of inorganic salts which have a definite crystalline structure. The formation of such bonds is based on Kossel's theory. Kossel (1916) postulated that the atomic structure of an inert gas represented a stable arrangement of electrons and that other atoms try to achieve that structure by losing or gaining the required number of electrons. An element preceding an inert gas in the periodic table is strongly electronegative and the element immediately following the inert gas is strongly electropositive. For example, chlorine is electronegative while sodium is electropositive. Electropositive elements may assume inert gas configuration by losing one electron as they have low ionization potential. Conversely electronegative elements may assume inert gas configuration by gaining one or more electrons, as they have stronger electron affinity. Thus sodium atom can lose one electron and acquire a positive charge, while chlorine atom can accept one electron and acquire a negative charge. Ionic bonds may be formed by the transfer of more than one electrons also. For example, one oxygen atom can form 'an ionic bond by accepting one electron from each of the two sodium atoms.

$$\begin{matrix} Na+O & \rightarrow & 2Na^{+} + O^{2-} \\ Na & & Na_2O \end{matrix}$$

Oxygen can also form ionic bonds by accepting two electrons from a single atom of the element which has a low second ionization potential. For example, one calcium atom can link to one oxygen atom by transferring its two electrons.

$$Ca{:}+O \rightarrow Ca^{2} + {:}O^{2-}$$

$$\searrow CaO$$

The number of the charge (+or-) or the number of electrons transferred during an ionic bond formation is called electrovalence number or the electrovalency of the participating atoms.

1.3.2. Covalent bonds:

Covalent bonds are the properties of non-metals. They have very high ionization potential and do not lose electrons easily and also of the larger organic molecules. During covalent bond formation there is no transfer of electrons; instead they are shared between the atoms. For example, two atoms of hydrogen are combined together to form hydrogen molecules (H2) oy sharing their single electron with each other.

$$H + H \rightarrow H\text{-}H\ (H_2)$$

Certain other gases such as nitrogen, fluorine, oxygen etc; also exist as diatomic molecules, where the two atoms are joined together with covalent bonds. In methane, 4 electrons of the carbon atom are shared with four electrons of four different hydrogen atoms.

H

:

H .. C ..H

The force of bond formation in covalent bonds is the same as that in ionic bonds i.e. electrostatic attraction between the two atoms, although this force of attraction develops in a different manner. When two atoms destined to link through a covalent bond, come within a definite range, the wave functions of the electrons of two atoms overlap each other. This overlapping causes the accumulation of negative charge between the two atomic nuclei. The accumulated negative charge in turn attracts the nuclei of the two atoms and thus they are held together (Fig. 2).

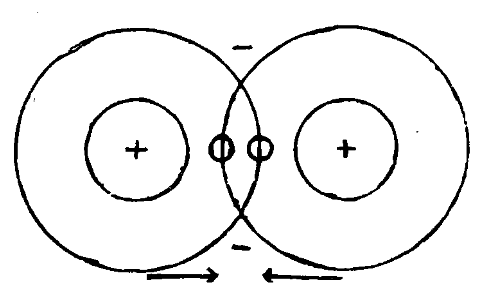


Fig. 2. Bond formation between two atoms

1.3.3. Co-ordinate covalent bonds:

Co-ordinate covalent bonds are also formed as a result of unequal sharing of electrons, as in polar covalent bonds. However, in this case the two electrons involved in the formation of a bond belong to the same atom. The atom which contributes the electrons is called donor and the atom receiving them is called acceptor. However, application of these terms (donor and acceptor) is rather erroneous, because there is no complete transfer of electrons. Formation of ammonium boron trifluoride complex involves the co-ordinate covalent bond. In this case the two unpaired electrons of nitrogen are involved in binding with boron. The co-ordinate covalent bonds are generally indicated with an arrow and the molecule in this case is also electrically

asymmetrical. Because of the shifting of the electrons towards one part, the other part of the molecule acquires a partial positive charge.

1.3.4. Hydrogen bonds:

When a hydrogen atom is covalently linked with a strongly electronegative atom such as oxygen or nitrogen, the bond develops a considerable ionic nature with the hydrogen forming the positive end of a dipole and the electronegative atom, the negative end. This positively charged hydrogen atom gets attracted towards the negative end of another molecule. This attraction weakly binds the two molecules. A number of molecules attracted to each other by these electrostatic forces may associate together to form large clusters of molecules with hydrogen acting as a bridge between the electronegative atoms. This attractive force which binds the hydrogen atom of one molecule with an electronegative atom of other molecules is known as the hydrogen bond or the proton bond. A proton donor group possesses positively charged hydrogen available for hydrogen bonding and the acceptor group has an unshared pair of electrons.

1.3.4.1 Types of hydrogen bonds

There are two types of hydrogen bonding, intermolecular and intramolecular. The intermolecular hydrogen bonding results in association of molecules and is influenced by the shape of the molecules. This type of H bonding changes the number, mass, shape and electronic structure of the participants in a system. Due to the intermolecular association the effective molecular weight of a structure is increased leading to higher melting and boiling points. Intramolecular hydrogen bonding takes place in molecules of the 3rd group; between the proton donor group and the proton acceptor group of the same molecule. This type of bonding occurs only when (1) the proton donor and one proton acceptor sites on the same molecules is in one proton acceptor sites on the same molecule is in a favourable spatial configuration, that is, the distance between the hydrogen of the donor group and the acceptor site is in between 1.4-2.5 A, (2) the angular orientation of the acceptor site does not differ much from the bond of the donor group A-H and (3) the molecules are in the cis or a gauche (skew) conformation(Fig.3). Intramolecular H bonding gives rise to chelation i.e. formation of ,6,7- heterocyclic rings. This reduces the solubilities of these molecules in water because these compounds cannot form H bonds with water



Gauche form Cis-form

Fig. 3 : Cis and gauche forms of the molecules.

1.3.4.2. Hydrogen bonds in biological systems

The fact that life originated in water ensures the prevalence of H bonds everywhere in the living system. The living cell contains 70-90% water. Almost all the biological molecules occur in aqueous environments; hence the role of the nearest neighbour water molecules is of great importance in the structure, conformation and arrangements of biopolymers. These biopolymers enter into intermolecular H bonding among themselves and with adjoining water molecules and almost all the polar molecules are interlinked through an extensive network of H bonded water molecules in a living system. The exhaustive intramolecular H bonding results in considerable folding in their skeletal structures of these macromolecules and produces microenvironments whose polarities range from non-polar .and highly polar depending on the number arrangements of water molecules in each micro region. The thermodynamic parameters (free energy, enthalpy and entropy) of the H bonds between the proton donor and acceptor side chain groups will vary with the polarity of their specific micro-environments. Though not directly, the extensive H bonding within a cell is responsible for the proper 'alignments of the non-polar molecules like lipids to form associations, the membranes etc. These are short range forces, operative for small distances only. They occur when two molecules come very close to each other. The existence of these forces was proposed by van der Waals. Van der Waals forces originate when two dipoles come together within short distances. Depending on the nature of dipoles three types of forces are recognized namely Keesom forces, Debye forces and London forces(Fig.4).

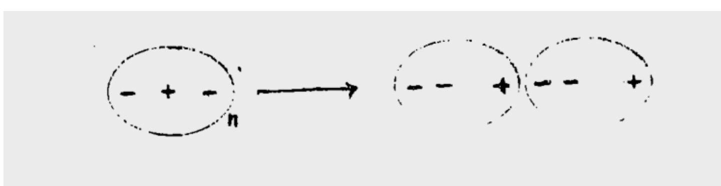
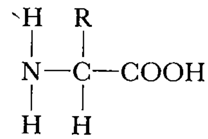


Fig 4 : Van der Walls Forces

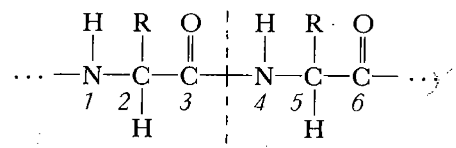
1.4 Polymerization of organic molecules :

The X rays diffracted from a single crystal interfere with one another in a manner which is determined solely by the position and electron density of the target atoms in the crystal. If the diffracted rays are allowed to fall upon a photographic plate, from the position and darkness of the spots on the plate, one can (at least in principle) locate the position and electron density of the diffracting atoms in the crystal. It was in 1951 that Pauling and Corey made the big break-through in our understanding of structure of proteins: they were able to determine from X-ray diffraction patterns that synthetic polypeptides formed of alpha amino acids all have a coiled,

helical form. In other words, the back-bone of the polypeptide chain coils around and around, to form a cylindrically shaped molecular helix. This can be easily understood now, in retrospect, as follows. Since all the alpha amino acids have the structural formula



and since these condense through the —CONH— linkage in the form



the atoms of the backbone of the chain, —N—C—C—, are repeated over and over again. The bonds can be beni around only so far, and, in the limit, carbon 6 falls almost directly above nitrogen 7, and the two are hydrogen bonded about 1.5 A apart. The diameter of the helix so formed is about 8 A. The helix has an open Qpre, about 2 A across,, and the R-groups, or side chains to the main structure, but out radially from the central axis of the cylinder.

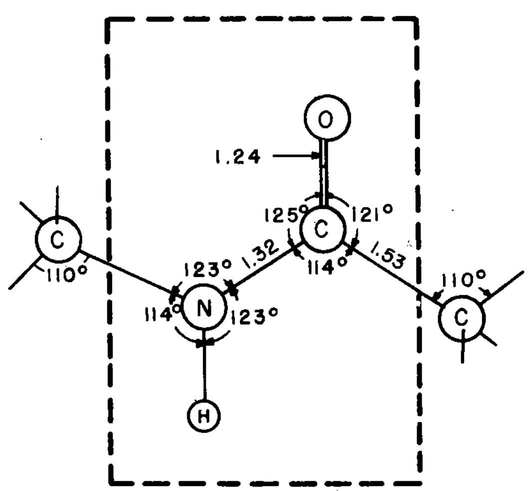


Figure 5. The Planar —CONH— Linkage (boxed)
between Amino Acids in a Protein. Lengths in angstroms.

Since the helical shape is a property of polyalpha amino acids, it was given the name "alpha-helix," and it is now probably the most famous structure of macromolecular physical chemistry. Figure 5 is a drawing, similar to the original disclosure, which shows the main chain (bold-bonds) and the positions of attached groups (—R); and which indicates the positions of the hydrogen bonds, the "bones" which give the helix rigidity.

It is now known to be the main structural component of a keratin-hair, wool, nail, muscle, and connective tissue, etc. Recently it has been traced in muscle to the contractile enzyme, myosin itself. One protein, of unquestioned importance, which has intrigued biological investigators for years, is hemoglobin, the "oxygen carrier" of the respiratory enzyme system, first crystallized and purified by Hoppe-Seyler in 1862. However, with a molecular weight of 67,000, its amino acid sequence and the physical structure of the molecule have only slowly yielded to persistent investigation.

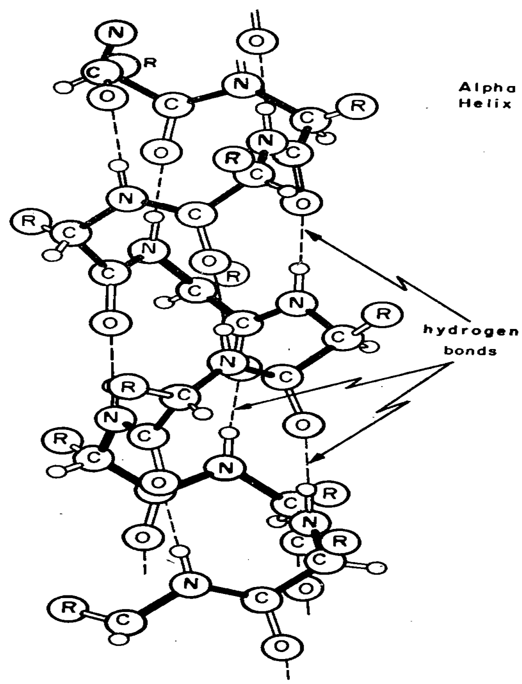


Figure 6. Schematic Representation of the Alpha Helix of Protein. Three complete turns are shown. They start at the bottom C, wind out toward the reader through the next —N—C—, then back in through the plane of the paper, etc. (After Pauling and Corey, 1953.)

The X-ray diffraction pattern of even single crystals was too formidable for analysis until M.'F. Perutz, about 1950, began to substitute heavy metals ions such as Hg+2 at particular spots on the molecule and to analyze the effects of these strong X-ray scatterers on the spectrum. With this technique, now known as the "method of isomorphous replacement," it was possible by 1960 to show the surprising result that the protein of the molecule at 6 angstroms' resolution looks like several intertwined worms, with the heme groups attached not a regular array at all. Studies continue on the amino acid sequence, and on the analysis of the X-ray diffraction pattern, in an effort to get even better resolution of the detailed structure of the hemoglobin molecule. Inherently simpler, myoglobin (one Fe+2 ion only) has yielded not only to 6 A analysis (1956) but even to 1.5 Arc-solution (1958), work for which Kendrew and his team received a Nobel

Prize in 1~61. The main features of this molecule are depicted in the drawing shown in Figure 7. The a-helix

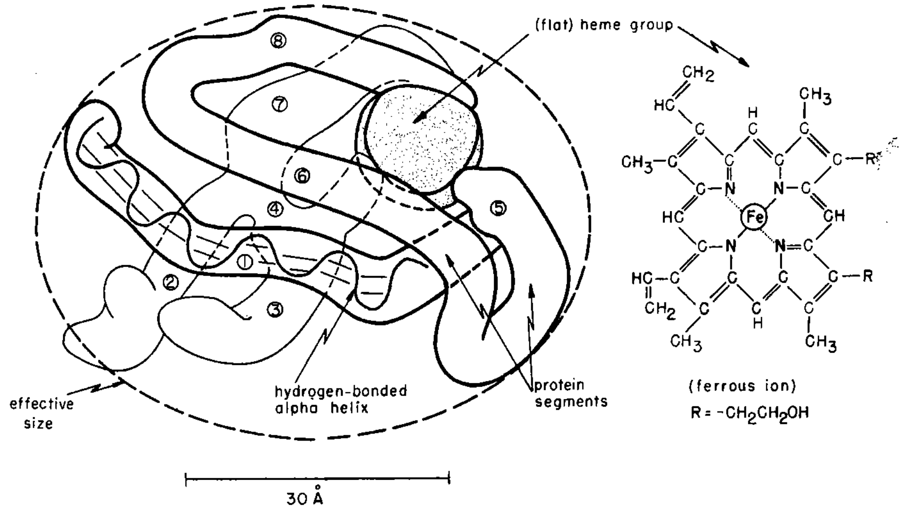


Figure 7. Molecule of Myoglobin. (Drawn from the Model of Kendrew, 1958.)

Hydrogen bond, forms the framework of the worm-like segments, sudden turns in which are thought to be' associated with the proline groups an amino acid residue of odd structural configuration. The heme group sits ex- posed, with the iron ion ready for oxidation or reduction, or, preferably, simply cornplexing with 0_2 picked up from air. Although this is the configuration of crystalline myoglobin, the shape of the molecule dissolved in salty water may be quite different for example; one can readily imagine the legs of this molecular octopus unfolding in the blood stream.

Structural knowledge of many other big molecules is rapidly becoming available. This is a subject of intense interest. Straight chains and helices) some coiled into balls, some folded back and forth to form rods, others with randomly coiled shapes, are known or imagined. These forms are illustrated in Figure 8.

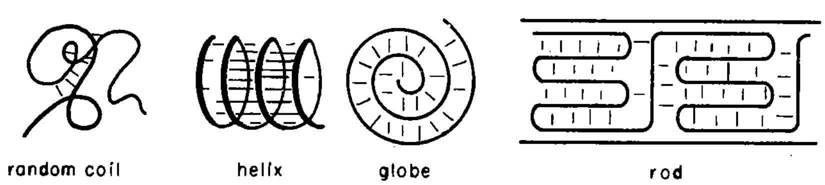


Figure 8. Some Molecular Shapes in Solution (schematic). Transitions one to another can be effected by change in pH, ionic composition, or temperature.

Receiving much attention in the hands of F.0.Schmitt and the MIT School has been collagen, the structural component of connective tissue, tendon, skin, cartilage, etc. (Figure 9). Formed of three interwound molecular helices of protein, with molecular dimensions approximately 3000 A long x 30 A in diameter, it cross-links end to end to form fibers, and then side to side to form either sheets (two dimensions) or blocks (three di- mensions) of connective

tissue with very varied physical properties: for ex- ample, tensile strength up to 100,000 Ibs/in2, equivalent to that of a steel wire of the same dimensions!

Now thought to be the basic information-carrier of the gene, and an ex- tremely important component of the nucleus of the cell, is desoxyribose- nucleic acid (DNA). At about 70 per cent relative humidity, it is an ex- tended, double-stranded helix, of molecular weight in the millions. Further discussion of the structure of DNA, and its sister nucleic acid, ribosenucleic acid (RNA), appears later in this chapter.

Now, the backbone of the helices of DNA and RNA is ribose, a sugar, polymerized through phosphate groups. Polymerized sugars are the second major structural component of living tissue cellulose and chitin are examples. Hyaluronic acid and glycogen are polysaccharides which take an integral part in the biochemistry of life. Thus glycogen 'is the form in which sugar is stored as an energy reserve in the liver, Polysaccharides, like proteins, take many forms in tissue. One which seems to be unique is the pleated sheet of cellulose.



Figure 9. Electron Micrograph of Collagen Fibers Carefully Lifted from Human Skin. Note how they are individually cross-segmented and collectively fused (Courtesy of J. Gross, Massachusetts General Hospital, and of Scientific American.)

Lipid molecules themselves are generally small, by comparison with the macromolecules discussed in this section . However they condense with proteins to form macromolecular lipoproteins, and with cellulose to form lipocelluloses, and thus also play a primary role in the structure of tissue.

Metal-organic molecules are varied and important in living tissue (Table 6-1). The bright light from the point of view of our knowledge is vitamin B12, a substituted cyanocobalt amide of

molecular weight 1357, used in treating pernicious anemia, growth failure in children, etc. The complete chemical composition was disclosed in 1955, and culminated with X-ray diffraction analysis of structure three years later.

STUDY QUESTIONS:

Short Questions

1. **What is the structure of an atom.**
2. **Mention any two types of Bonds.**
3. **What is Covalent Bond.**
4. **Define polymerization.**

Long Questions

1. **Give an account formation of molecules from atoms?**
2. **Describe the various chemical bonds.**
3. **Write about the role of X-ray diffraction studies in Polymerization?**

Unit II

Structure

2.1 Laws of thermodynamics:

The statements of the Three Laws are given below ;

The First Law : The first law states simply that energy can be transformedFrom one form to another but cannot be created or destroyed. After the equivalence of matter and energy were recognized (and proved in nuclear reactions), the law was generalized still further to read: "mass-energy" in- stead of "energy." The Law stands as written, needing no extension, for all cases in which any form of energy is converted into heat: 100 per cent conversion can always be realized.

The Second Law: For any machine which converts heat into mechanical work, chemical into electrical energy, or the like, it is a universal experience that, only attraction can be converted; the rest remains unavailable and unconverted. There is .thus an amount of unavailable energy as well as available energy from the conversion. The unavailable, it would be logical to assume, is the heat energy which must remain in the molecules of which the final state (i.e., the product) is composed.

The Third Law: At 0°K (-273.16°C), the absolute zero of temperature, at which all molecular motion has ceased, matter should be in a state of perfect order, the molecules being perfectly aligned or oriented, and perfectly quiet. This law is concerned with the absolute heat energy contained in molecules at any temperature. Although our present interest is in changes from one state to another, rather than absolute quantities in any state, the absolute quantities disclosed via the Third Law permit easy evaluation of the changes.

2.2 Principles and Applications:

The internal energy of a body is defined as the sum total of all the kinetic and potential energy contained within the body, When expressed per gram molecular weight it is given the symbol U cal/mole, and is a "state vari- able," that is, one whose value depends only upon the temperature, pres- sure, and composition, irrespective of how it arrived at this condition. Heat energy, (that contained in the motion of the molecules), potential energy of the electron cloud of the atom, and the binding energy of the nucleus all contribute to the internal energy.

If a transformation takes place in one molecular weight of a substance, two things in general can occur: energy can be taken in by the substance, and work can be done. If an amount of energy, q, is taken in, and an amount of work, w, is done, the difference, q — w, must be the increase in energy of the substance during the process; this difference must be stored as internal energy, and hence the change in internal energy is:

Now hU =q— w is the concise, algebraic statement of the First Law. The concepts are illustrated in Figure 10.

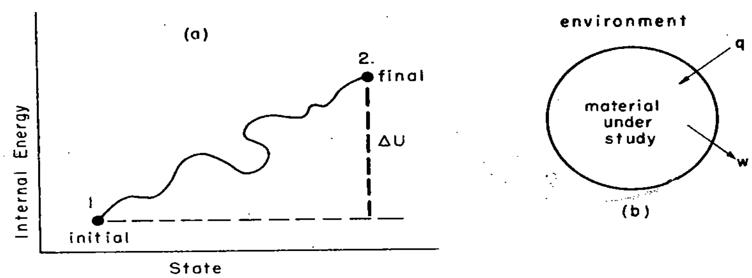


Figure.10 :The First Law of Thermodynamics: (a) a state diagram showing internal energy change. All, during a process; (b) the process: heat taken in, q, and work done, w.

More Detailed Consideration of the First Law.

One could generalize to complex, non molar quantities of varied composition; the law would still be conceptually the same.The first law can be extended into a more useful form for processes taking place at constant pressure. Since any substance, this book, for example, has an individual and independent existence in space, and since it occupies a certain volume and has an area upon which the air pressure (i.e., weight of the column of air above it) is 15 lb/sq in., the book does not have as much internal energy as it would have if it were in a vacuum, because it already has done a considerable amount of work against atmospheric pressure. That is, it has already expended enough energy (or "work of expansion"), W, to rollback the atmosphere and create a hole or vacuum in which it can exist. Hence the internal energy

$$U = KE+PE-W$$

The work of expansion, W, can be easily evaluated. Consider the cylinder with frictionless piston of area, A, enclosing a volume of gas, V. From the definition of work:

Work = force x distance

$$= PA \times AV/A$$

$$= a\ V=P(V - V$$

Since we are considering an initial state, V\, of zero volume, in general W = PV. Substituting,

$$U = KE + PE - PV$$

$$= H - PV$$

Where H is the internal energy contained per mole in a vacuum (when P = 0). The quantity, H, is called heat content, or preferably enthalpy because really potential energy as well as heat kinetic energy is included. A little thought about the definition will lead one to the conclusion that H should be a very useful quantity for comparison purposes because its value is independent of any volume change which may accompany a transformation or process. Further, for the case of chemical reactions, $\Delta H = H_2 - H_1$ (note the parallel with ΔU) must be identical with q, the heat taken in during the process for the case in which the only work done is that of expansion; i.e., q = ΔH. Many biological processes occur in solution, with no appreciable change in volume, and in these cases ΔU = ΔH.

Now ΔH = q may be positive or negative depending upon which is larger, the enthalpy of the final or of the initial state. The former characterizes an endothermic reaction; the later an exothermic reaction. As a general rule anabolic reactions are endothermic; catabolic reactions are exothermic. More specifically, the synthesis of proteins in the metabolism of the living system is endothermic; the combustion of glycogen and other food stores is exothermic. For chemical reactions the value of *q* or ΔH, the "heat of reaction," can be measured calorimetrically, and quite accurate values obtained. For in- stance, for the simplest reaction

$H_2 + 1/2\ O_2 = H_2O$

the heat of reaction

$\Delta H = H_{final} – H_{initial}$

$= H(l\ mole\ H_2O) - H(1molecH_2 + 1/2\ mole\ 0_2)$

and although the absolute value of the enthalpy (or internal energy) for neither reactants nor product is known (Who knows how to determine the sum of all the potential energies in the nucleus, for example?), the difference, AH, can be obtained with great precision: —57,798 cal/mole at 25°C, the minus sign indicating that the reaction is exothermic.

An especially useful heat reaction is the heat of formation, AHf, the enthalpy change which occurs during the reaction by which the molecule of interest is formed from its elements. Actually the example above was a formation reaction. Another now follows:

$6C(s) + 6H_2\ (g) + 3\ O_2\ (g) = C_6H_{12}O_6$ (glucose)

ΔH = -279,800 cal/mole

From a table of heats of formation, heats of reaction can be computed as

$\Delta H = (\Delta H_1)$products- (ΔH_1)reactions

The heat of combustion or burning of glucose could be computed, from heats of formation, from the following reaction:

$C_6H_{12}O_6 + 60_2(g) = 6H,O(1) + 6(CO_2(g)$ ΔH = -669,580 cal/mole

The fuel value of foods is usually expressed in units of thousands of calories: i.e., kilocalories (kcal), kilogram calories (kg cal), or Calories (Cal). Hence the fuel value of glucose is 669.58 kcal/mole. The quantity of heat given off by living animals can be measured either calorimetrically or by the CO_2 produced, (The two measurements agree!), and when measured under conditions of a carefully defined rest, give a value related to the internal work required to keep the living system alive. This basal metabolic rate is about 70 kcal/hr, (about 1400 kcal/day) for a normal man. In other units, the basal metabolic rate amounts to about 0.1 horsepower (hp) continuously.

2.2.1 Free Energy and Entropy :

The Second Law of Thermodynamics does not violate the first, but rather extends it. It says: Whenever energy is transformed from one kind into another, only a fraction of the internal energy (enthalpy, if pressure is constant) change is available for doing useful work; the rest remains as heat energy of the mole- cules left at the completion of the reaction,,. Corollaries, although seemingly unrelated, are the following: heat energy always passes from the hot to the cold body; water always runs downhill; if energy available for doing work can decrease during the course of a process, the process will proceed spontaneously, although not necessarily at a fast rate. (That last phrase is a very important one!)

In algebraic terms, the Second Law can be expressed as:

$$\Delta H = \Delta F + Q$$

Here ΔF is the maximum available work, the "free" energy, which can be extracted from ΔH, and Q is the unavailable energy.. Note that both ΔF and Q as does ΔH, have units kcal/mole (i.e., Cal/mole). The word "maximum" needs amplification. It is a fact of common experience that any mechanical job can be done in several ways, some ways more efficient than others. If the job is done by the hypothetical frictionless machine, with minimum loss of energy, it is then done the most efficiently. By analogy, work can be extracted from a process in many ways, some more efficient than others. The hypothetical conditions of no waste are given .the special name, reversible conditions; AF is therefore the maximum work avail- able under reversible conditions. One practical system from which nearly maximum work can be extracted is the electrochemical one, a battery for example; or, more pertinent here, the concentration cells which exist and deliver energy at living membranes.

2.3 Bioenergetics :

Possible routes for the evolution of cell energetics are considered. It is assumed that u.v. light was the primary energy source for the precursors of the primordial living cell and that primitive energetics might have been based on the use of the adenine moiety of ADP as the u.v. chromophore. It is proposed that the excitation of the adenine residue facilitated phosphorylation of its amino group with subsequent transfer of a phosphoryl group to the terminal phosphate of ADP to form ATP. ATP-driven carbohydrate synthesis is considered as a mechanism for storing u.v.-derived energy, which was then used in the dark. Glycolysis presumably produced compounds like ethanol and CO2, which easily penetrate the membrane and therefore were lost by the cell. Later lactate-producing glycolysis appeared, the end product being non-penetrant

and, hence, retained inside the cell to be utilized to regenerate carbohydrates when light energy became available. Production of lactate was accompanied by accumulation of equimolar H+. To avoid acidification of the cell interior, an F0-type H+ channel was employed. Later it was supplemented with F1. This allowed the ATP energy to be used for 'uphill' H+ pumping to the medium, which was acidified due to glycolytic activity of the cells. In the subsequent course of evolution, u.v. light was replaced by visible light, which has lower energy but is less dangerous for the cell. It is assumed that bacteriorhodopsin, a simple and very stable light-driven H+ pump which still exists in halophilic and thermophilic Archaea, was the primary system utilizing visible light. The delta mu-H+ formed was used to reverse the H(+)-ATPase, which began to function as H(+)-ATP-synthase. Later, bacteriorhodopsin photosynthesis was substituted by a more efficient chlorophyll photosynthesis, producing not only ATP, but also carbohydrates. O2, a side product of this process, was consumed by the H(+)-motive respiratory chain to form delta mu-H+ in the dark. At the next stage of evolution, a parallel energy-transducing mechanism appeared which employed Na+ instead of H+ as the coupling ion

ATP can easily release and store energy by breaking and re-forming the bonds between its phosphate groups. This characteristic of ATP makes it exceptionally useful as a basic energy source for all cells. - In the process of photosynthesis, plants convert the energy of sunlight into chemical energy stored in the bonds of carbohydrates. Photosynthetic organisms capture energy from sunlight with pigments. An electron carrier is a compound that can accept a pair of high energy electrons and transfer them, along with most of their energy, to another molecule. Photosynthesis uses the energy of sunlight to convert water and carbon dioxide into high energy sugars and oxygen. Among the most important factors that affect photosynthesis are temperature, light intensity, and the availability of water. - Organisms get the energy they need from food. Cellular respiration is the process that releases energy from food in the presence of oxygen. Photosynthesis removes carbon dioxide from the atmosphere and cellular respiration puts it back. Photosynthesis releases oxygen into the atmosphere, and cellular respiration uses that oxygen to release energy from food. In the absence of oxygen, fermentation releases energy from food molecules by producing ATP. For short, quick bursts of energy, the body uses ATP already in muscles as well as ATP made by lactic acid fermentation. For exercise longer than about 90 seconds, cellular respiration is the only way to continue generating a supply of ATP. Vocabulary: ATP ADP autotroph heterotroph Photosynthesis pigment chlorophyll chloroplast Thylakoid stroma NADP+/NADPH calorie Cellular respiration aerobic anaerobic fermentation Glucose.

2.4. Coupling of chemical reactions :

A cell can be thought of as a small, bustling town. Carrier proteins move substances into and out of the cell, motor proteins carry cargoes along microtubule tracks, and metabolic enzymes busily break down and build up macromolecules. Even if they would not be energetically favorable (energy-releasing, or exergonic) in isolation, these processes will continue merrily along if there is energy available to power them (much as business will continue to be done in a town as long as there is money flowing in). However, if the energy runs out, the reactions will grind to a halt, and the cell will begin to die. Energetically unfavorable reactions are “paid for” by linked, energetically favorable reactions that release energy. Often, the "payment" reaction involves one particular small molecule: adenosine triphosphate, or ATP.

How is the energy released by ATP hydrolysis used to power other reactions in a cell? In most cases, cells use a strategy called **reaction coupling**, in which an energetically favorable reaction

(like ATP hydrolysis) is directly linked with an energetically unfavorable (endergonic) reaction. The linking often happens through a **shared intermediate**, meaning that a product of one reaction is “picked up” and used as a reactant in the second reaction. When two reactions are coupled, they can be added together to give an overall reaction, and the Δ*G* of this reaction will be the sum of the Δ*G* values of the individual reactions. As long as the overall Δ*G* is negative, both reactions can take place. Even a very endergonic reaction can occur if it is paired with a very exergonic one (such as hydrolysis of ATP). For instance, we can add up a pair of generic reactions coupled by a shared intermediate, B, as follows^22start superscript, 2, end superscript:

A+BA⇋B⇋C+D⇋C+DΔ*G*=*X*Δ*G*=*Y*Δ*G*=*X*+*Y*

You might notice that the intermediate, B, doesn't appear in the overall coupled reaction. This is because it appears as a both a product and a reactant, so two Bs cancel each other out when the reactions are added. When reaction coupling involves ATP, the shared intermediate is often a phosphorylated molecule (a molecule to which one of the phosphate groups of ATP has been attached). As an example of how this works (Fig.11), let’s look at the formation of sucrose, or table sugar, from glucose and fructose{3,4}3,4start superscript, 3, comma, 4, end superscript.

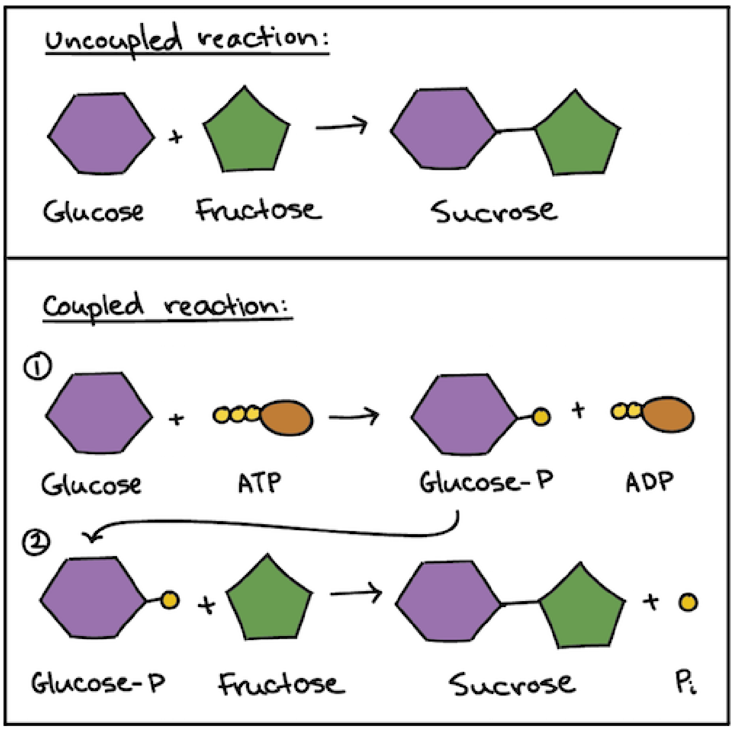


Fig.11. Illustration of reaction coupling using ATP.

In the uncoupled reaction, glucose and fructose combine to form sucrose. This reaction is thermodynamically unfavorable (requires energy).

2.5. Redox potential:

The Redox potentials are used to characterize the free energy cost and direction of reactions involving electron transfer, one of the most ubiquitous and important of biochemical reactions. Such reduction-oxidation reactions are characterized by a free energy change that shares some conceptual features with that used to describe pKa in acid-base reactions where

proton transfer is involved rather than electron transfer. In this vignette, one of the most abstract in the book, we discuss how the redox potential can be used as a measure of the driving force for a given oxidation-reduction reaction of interest. By way of contrast, unlike the pH, there is no sense in which one can assign a single redox potential to an entire cell.

The redox potential, or more accurately the reduction potential, of a compound refers to its tendency to acquire electrons and thereby to be reduced. Some readers might remember the mnemonic "OILRIG" which reminds us that "oxidation is loss, reduction is gain", where the loss and gain are of electrons. Consider a reaction that involves an electron transfer: $A_{ox} + ne^{-} \leftrightarrow A_{red}$ where n electrons are taken up by the oxidized form (A_{ox}) to give the reduced form (A_{red}) of compound A. The redox potential difference ΔE between the electron donor and acceptor is related to the associated free energy change ΔG of the reaction via $\Delta G = nF\Delta E$ where n is the number of electrons transferred and F is Faraday's constant (96,485 J/mol/V or ≈100 kJ/mol/V). By inspecting tabulated values of these potentials, it is possible to develop an intuition for the tendency for electron transfer and hence, of the direction of the reaction.

Though ATP is often claimed to be the energy currency of the cell, in fact, for the energetic balance of the cell the carriers of reducing power are themselves no less important. The most important example of these carriers is the molecule NADH in its reduced or oxidized (NAD^{+}) forms. We can use the redox potential to connect these two molecular protagonists, and estimate an upper bound on the number of ATP molecules that can be produced from the oxidation of NADH (produced, for example, in the TCA cycle). The NAD^{+}/NADH pair has a redox potential of E = -0.32 V and it is oxidized by oxygen to give water (protons coming from the media) with a redox potential of E = +0.82 V. Both are shown in Figure 1 as part of a "redox tower" of key biological half reactions that can be linked to find the overall redox potential change and thus the free energy. For the reaction considered above of NADH oxidation by oxygen, the maximal associated free energy that can be extracted is thus

$$\Delta G = n \times F \times \Delta E = 2 \times 100 \text{ kJ/(mol} \times \text{V)} \times (0.82-(-0.32)) \text{ V} = 230 \text{ kJ/mol} \approx 90\ k_BT,$$

where n=2 and F≈100kJ/mol/V. As ATP hydrolysis has a free energy change of ≈50kJ/mol under physiological conditions we find that 228 kJ/mol suffices to produce a maximum of 228/50≈4.5 ATPs. In the cell, oxidation of NADH proceeds through several steps in respiration and results in the transfer of 10 protons across the membrane against the electro-chemical potential These proton transfers correspond to yet another way of capturing biochemical energy. This energy is then used by the ATPase to produce 2-3 ATPs. We thus find that about half of the energy that was released in the transfer of electrons from NADH to oxygen is conserved in ATP. Ensuring that the reaction proceeds in a directional manner to produce ATP rather than consume it requires that some of the energy is "wasted" as the system must be out of equilibrium.

Why should one discuss redox potentials of half reactions and not free energies of full reactions? The units themselves owe their origins to the ability in the field of electrochemistry to measure in the lab the voltage difference, i.e. the potential measured in volts, across two chambers that contain different electron carriers, and to stop the net reaction with a voltage. The usefulness of redox potentials for half reactions lies in the ability to assemble combinations of different donors and acceptors to assess the thermodynamic feasibility and energy gain of every considered reaction. If you have k possible electron transfer compounds, the $\sim k^2$ possible reactions can be predicted based on only the k redox potentials.

Just as we speak of the pH of a solution, at first guess, we might imagine that it would be possible to speak of an apparently analogous redox potential of the cell. Knowing the concentration of the reduced and oxidized forms of a given reaction pair defines their pool redox potential via the relation

$$E = E_0 - \frac{RT}{nF} ln \frac{[A_{red}]}{[A_{ox}]}$$

This equation (a so-called Nernst equation) provides the value of the redox potential under concentration conditions typical of the cell as opposed to the standard state conditions (where by definition $[A_{red}]=[A_{ox}]$). As an example, consider the donation of an electron to NAD^+ resulting in the oxidized form NADH. In the mitochondrial matrix a ratio of 10-fold more of the oxidized form is reported as shown in Table 1. In this case, we find the factor is ≈30 mV and thus the redox potential changes from -0.32 V to -0.29 V. To make sure the direction of effect we got is sensible we notice that with an overabundance of the oxidized form the tendency to be oxidized by oxygen is somewhat lower as seen by the fact that the redox potential is now closer than before to that of the oxygen/water electron exchanging pair (+0.82V)**.**

A cell is not at equilibrium and there is weak coupling between different redox pairs. This situation leads to the establishment of different redox potentials for coexisting redox pairs in the cell. If the fluxes of production and utilization of the reduced and oxidized forms of a redox pair, A_{red} and A_{ox} and another B_{red} and B_{ox}, are much larger than their interconversion flux, $A_{red}+B_{ox}$<–>$A_{ox}+B_{red}$ then A and B can have very different redox potentials. As a result it is ill defined to ask about the overall redox potential of the cell as it will be different for different components within the cell. By way of contrast, the pH of the cell (or of some compartment in it) is much better defined since water serves as the universal medium that couples the different acid-base reactions and equilibrates what is known as the chemical potential of all species.

For a given redox pair in a given cell compartment the concentration ratio of the two forms prescribes the redox potential in a well-defined manner. Compounds that exchange electrons quickly will be in relative equilibrium and thus share a similar redox potential. To see how these ideas play out, it is thus most useful to consider a redox pair that partakes in many key cellular reactions and, as a result, is tightly related to the redox state of many compounds. Glutathione in the cytoplasm is such a compound as it takes part in the reduction and oxidation of the highly prevalent thiol bonds (those containing sulfur) in cysteine amino acids of many proteins. Glutathione is a tripeptide (composed of 3 amino acids), the central one a cysteine which can be in a reduced (GSH) or oxidized form where it forms a dimer with a cysteine from another glutathione molecule (denoted GSSG). The half reaction for glutathione is thus 2 x GSH <-> GSSG + $2e^-$ + $2H^+$. The other half reaction is often a sulfur bond that is “opened up” in a receptive protein thus being kept in the reduced form owing to the constant action of glutathione. Glutathione is also a dominant player in neutralizing reactive compounds that have a high tendency to snatch electrons and thus oxidize other molecules. Such compounds are made under oxidative stress as for example when the capacity of the electron transfer reactions of respiration or photosynthesis is reached. Collectively called ROS (reactive oxygen species) they can create havoc in the cell and are implicated in many processes of aging. The dual role of glutathione in keeping proteins folded properly and limiting ROS as well as its relatively high concentration and electron transfer reactivity make it the prime proxy for the redox state of the cell. The

concentration of glutathione in the cell is ≈10mM making it the second most abundant metabolite in the cell (after glutamate) ensuring that it plays a dominant role as an electron donor in redox control of protein function. In other functions of cells there are other dominant electron pairs. In biosynthetic anabolic reactions the $NADP^+$/NADPH pair and in breakdown catabolic reactions it is NAD+/NADH. How does one go about measuring redox potentials in living cells? Yet another beneficiary of the fluorescent protein revolution was the subject of redox potentials. A reporter GFP was engineered to be redox sensitive by incorporation of cysteine amino acids that affect the fluorescence based on their reduction by the glutathione pool. Figure 2 shows the result of using such a reporter to look at the glutathione redox potential in different compartments of a diatom.

From measurements of the redox state of the glutathione pool in different cellular organelles and under varying conditions we can infer the ratio of concentrations of the reduced to oxidized forms. Values range from about -170 mV in the ER and in apoptotic cells to about -300 mV in most other organelles and in proliferation cells. Given that the standard redox potential of glutathione is -240 mV. Using the Nernst equation (or equivalently, from the Boltzmann distribution), a ten-fold change in the product/reactant ratio corresponds to an increase of ≈ 6 kJ/mol in free energy (≈2 k_BT). Given the 2 electrons transferred in the GSH/GSSG reaction this concentration ratio change is usually equal to 30mV, though for glutathione, the stoichiometry of 2 GSH molecules merging to one GSSG covalently-bound molecule makes this only an approximation. The 100 mV change reported across conditions reflects a ratio of concentrations between about equal amounts of the reduced and oxidized forms (in apoptotic cells) to over 1,000 fold more concentration of the reduced form. Indeed in most cellular conditions the oxidized form is only a very small fraction of the overall pool but still with physiological implications.

2.6. NADP/NADPH:

The Nicotinamide adenine dinucleotide phosphate (NADP+) is an enzymatic cofactor involved in metabolic redox and cell signaling reactions. Its main function in animal metabolism is to shift electrons from one redox reaction to another. During these reactions, the coenzyme cycles between the electron donating reduced form (NADPH) and the electron accepting oxidized form (NADP+). The major source of NADPH produced in animal cells is the oxidative branch of the pentose phosphate pathway (PPP). NADPH is involved with macromolecule biosynthesis by providing reducing power. These characteristics have made NADPH an important molecule in cancer cell proliferation and metabolism. NADPH is also involved in the accumulation of Reactive Oxygen Species (ROS) and protecting against its toxicity. It is also involved in anabolic pathways, such as fatty acid elongation and lipid and cholesterol synthesis. Understanding the metabolism of these cofactors has become important to developing new therapeutic methods against pathological disease states. Introduction Cell Biolabs' NADP+ /NADPH Assay Kit is a simple colorimetric assay that can measure both NADP+ and NADPH present in biological samples such as cell lysates or tissue extracts in a 96-well microtiter plate format. The kit is specific for NADP+ , NADPH, and their ratio. The kit will not detect NAD or NADH. Each kit provides sufficient reagents to perform up to 100 assays, including blanks, NADP+ standards and unknown samples. The total NADP+ /NADPH concentrations of unknown samples are determined by comparison with a known NADP+ standard. Determination of both NADP+ and NADPH requires two separate samples for quantification. NADP+ and NADPH do not need to

be purified from samples, but rather can be extracted individually with a simple acid or base treatment prior to performing the assay. The kit has a detection sensitivity limit of approximately 4 nM NADP+(Fig. 12).

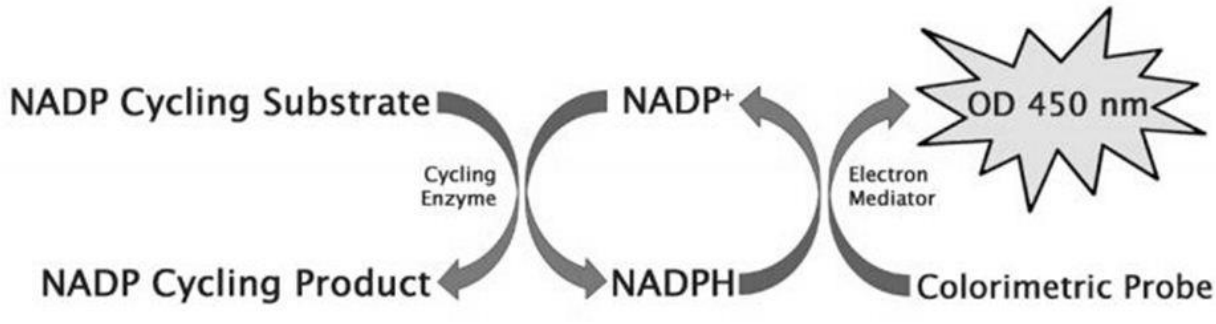


Fig. 12 :$NADP^+$/NADPH Cycling assay principle.

NADPH Extraction Procedure: To measure NADPH and destroy NADP+ , add 25 µL of sample to a microcentrifuge tube. Add 5 µL of 0.1 N NaOH and mix thoroughly. Incubate the tube at 80°C for 60 minutes and protected from light. Centrifuge the tube to pool all sample solution. Add 20 µL of 1X Assay Buffer to shift the pH of the sample back to neutral. Vortex to mix and centrifuge to pool sample. Sample pH should be between 6.0 and 8.0; if not, neutralize accordingly with acid or base. Keep sample on ice until assaying. NADP+ Extraction Procedure: To measure NADP+ and destroy NADPH, add 25 µL of sample to a microcentrifuge tube. Add 5 µL of 0.1 N HCl and mix thoroughly. Incubate the tube at 80°C for 60 minutes and protected from light. Centrifuge the tube to pool all sample solution. Add 20 µL of 1X Assay Buffer to shift the pH of the sample back to neutral. Vortex to mix and centrifuge to pool sample. Sample pH should be between 6.0 and 8.0; if not, neutralize accordingly with acid or base. Keep sample on ice until assaying.

2.7. Free Energy:

2.7.1 Free Energy from electromagnetic waves :

Light or electromagnetic radiation is a form of energy that is transmitted through space at a constant velocity of 5 x 10^8 ms^{-1} . These radiations are said to have dual nature exhibiting both wave and particle characteristics. The dual character is indeed useful for understanding the interactions of radiations with matter.

2.7.2 Wave Theory of Electromagnetic Radiation :

According to this theory, the electromagnetic radiations travel in the form of waves. This wave motion consists of oscillating electric and magnetic Fields directed perpendicular to each other and perpendicular to the direction of propagation of the wave as shown in the Figure 13.

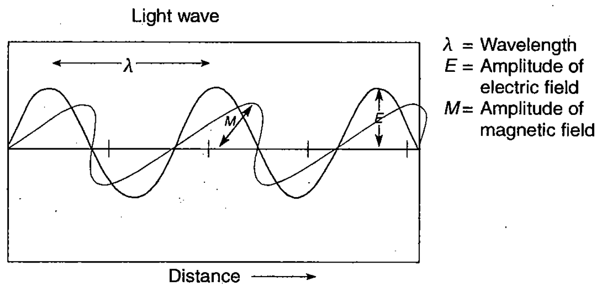


Fig.13 :The electric and magnetic components of electromagnetic radiation

In spectroscopic studies, the effects associated with the electrical component of the electromagnetic wave are important. The propagation of vibrations in the electrical Field only is shown in the Figure 14.

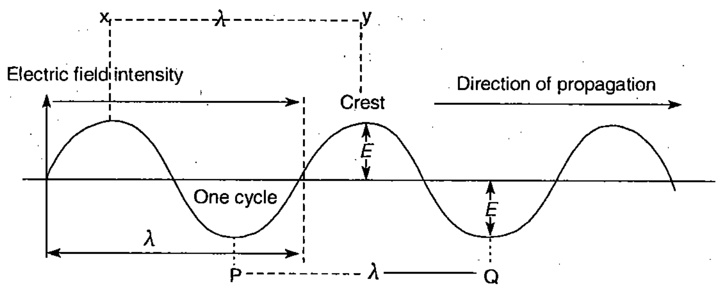


Fig.14: Propagation of vibration in the electricfield of electromagnetic radiation

The points X, Y, P and 0 on the wave represent the maximum disturbances in the electric field. The distance from the mean position is known as the amplitude of the wave. The distance from the crest X to crest Y (or from valley P to valley Q) is the wavelength X. The number of complete wavelength units passing through a given point per second is called frequency v. These two quantities are related to each other, and given where c is the velocity of the electromagnetic wave. Since c is constant ($3 \times 10^{2} ms^{-1}$ in vacuum) for all types of electromagnetic radiations, the above relation may be expressed as v x 1/λ

Reciprocal of wavelength, i.e., 1/λ is called wave number, ν Hence equation I may be written as, The wavelength is expressed in terms of centimetre (cm), metre (m), micron (~) or micrometer OLm) or angstrom (A) units. The other commonly used unit is nanometer (nm) where I nm $=10^{9}$. Frequency is measured as cycles per second (cps) called Hertz (Hz) or kilocycles per second (kHz) or megacycles per second (MHz). The wave number is the number of waves per unit distance and is expressed in the units of cm^{-1} called Kaysers (K). Sometimes kilokayser (kK) is also used.

STUDY QUESTIONS:

Short Questions

1. **State the First Law of thermodynamics.**
2. **What is Bioenergetics?**
3. **Define Redox Potential.**

4. **What is Free Energy**

Long Questions

1. **Explain the Principle and Application of first two laws of thermodynamics.**
2. **Describe the coupling of chemical reactions**
3. **Write about role of NADP and NAPH in Biological Systems?**

UNIT III

3.1 Natural Radiation :

Natural Rediation deals about the properties of natural light with emphasis on Photoelectric Light and Photodynamic sensitization. An understanding of LASER and its applications .Then acquiring knowledge on effect of radiations on macromolecules and the delayed effects of radiation. Further knowing the measurement of radio activity using Geiger Muller counter and Isotopes as tracers. Finally understanding the technique of Autoradigraphy and its applications.

When a surface is exposed to electromagnetic radiation above a certain threshold frequency (typically visible light for alkali metals, near ultraviolet for other metals, and extreme ultraviolet for non-metals), the radiation is absorbed and electrons are emitted. This phenomenon was first observed by Heinrich Hertz in 1887. Johann Elster (1854-1920) and Hans Geisel (1855-1923), students in Heidelberg, developed the first practical photoelectric cells that could be used to measure the intensity of light. In 1902, Philipp Eduard Anton von Lenard observed that the energy of individual emitted electrons increased with the frequency (which is related to the color) of the light. This appeared to be at odds with James Clerk Maxwell's wave theory of light, which was thought to predict that the electron energy would be proportional to the intensity of the radiation.In 1905, Albert Einstein solved this apparent paradox by describing light as composed of discrete quanta, now called photons, rather than continuous waves. Based upon Max Planck's theory of black-body radiation, Einstein theorized that the energy in each quantum

of light was equal to the frequency multiplied by a constant, later called Planck's constant. A photon above a threshold frequency has the required energy to eject a single electron, creating the observed effect. This discovery led to the quantum revolution in physics and earned Einstein the Nobel Prize in Physics in 1921.

Natural Rediation deals about the properties of natural light with emphasis on Photoelectric Light and Photodynamic sensitization. An understanding of LASER and its applications .Then acquiring knowledge on effect of radiations on macromolecules and the delayed effects of radiation. Further knowing the measurement of radio activity using Geiger Muller counter and Isotopes as tracers. Finally understanding the technique of Autoradigraphy and its applications.

3.2 Properties of light:

Visible light consists of electromagnetic waves that behave like other waves. Hence, many of the **properties of light** that are relevant to microscopy can be understood in terms of light’s behavior as a wave. An important property of light waves is the **wavelength**, or the distance between one peak of a wave and the next peak. The height of each peak (or depth of each trough) is called the **amplitude**. In contrast, the **frequency** of the wave is the rate of vibration of the wave, or the number of wavelengths within a specified time period (Figure 13).

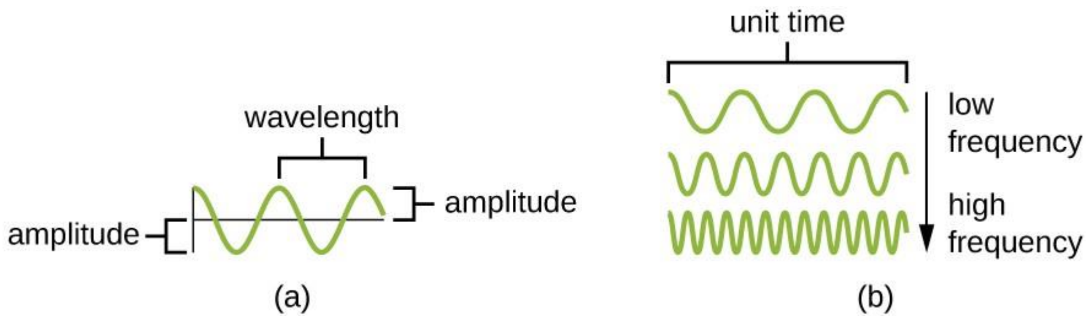


Fig.15 Amplitude and Frequency

3.2.1 Interactions of Light :

Light waves can also interact with each other by **interference**, creating complex patterns of motion. Dropping two pebbles into a puddle causes the waves on the puddle’s surface to interact, creating complex interference patterns. Light waves can interact in the same way. In addition to interfering with each other, light waves can also interact with small objects or openings by bending or scattering. This is called **diffraction.** Diffraction is larger when the object is smaller relative to the wavelength of the light (the distance between two consecutive peaks of a light wave). Often, when waves diffract in different directions around an obstacle or opening, they will interfere with each other.

3.2.2 Lenses and Refraction :

In the context of microscopy, **refraction** is perhaps the most important behavior exhibited by light waves. Refraction occurs when light waves change direction as they enter a

new medium (Figure 3). Different transparent materials transmit light at different speeds; thus, light can change speed when passing from one material to another. This change in speed usually also causes a change in direction (refraction), with the degree of change dependent on the angle of the incoming light.

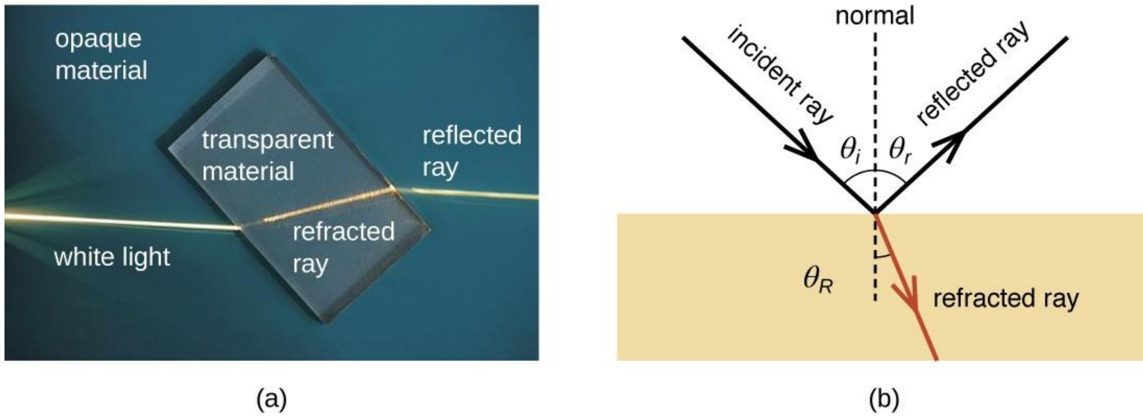


Figure 16. (a) Refraction occurs when light passes from one medium, such as air, to another, such as glass, changing the direction of the light rays. (b) As shown in this diagram, light rays passing from one medium to another may be either refracted or reflected.

The extent to which a material slows transmission speed relative to empty space is called the **refractive index** of that material. Large differences between the refractive indices of two materials will result in a large amount of refraction when light passes from one material to the other. For example, light moves much more slowly through water than through air, so light entering water from air can change direction greatly. We say that the water has a higher refractive index than air

When light crosses a boundary into a material with a higher refractive index, its direction turns to be closer to perpendicular to the boundary (i.e., more toward a normal to that boundary; see Figure 4). This is the principle behind lenses. We can think of a lens as an object with a curved boundary (or a collection of prisms) that collects all of the light that strikes it and refracts it so that it all meets at a single point called the image point (focus). A convex lens can be used to magnify because it can focus at closer range than the human eye, producing a larger image. Concave lenses and mirrors can also be used in microscopes to redirect the light path. Figure 5 shows the focal point (the image point when light entering the lens is parallel) and the focal length (the distance to the focal point) for convex and concave lenses.

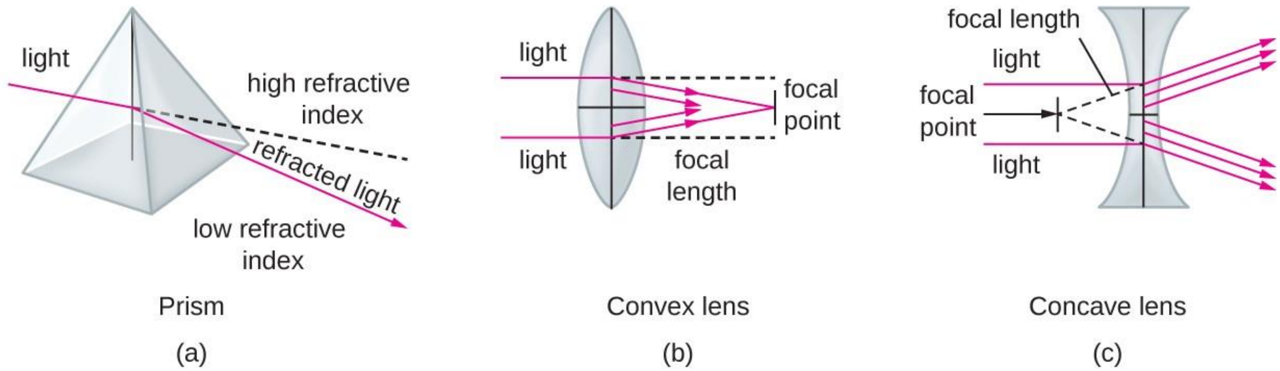


Fig 17 The human eye contains a lens that enables us to see images. This lens focuses the light reflecting off of objects in front of the eye onto the surface of the retina, which is like a screen in the back of the eye. Artificial lenses placed in front of the eye (contact lenses, glasses, or microscopic lenses) focus light before it is focused (again) by the lens of the eye, manipulating the image that ends up on the ` (e.g., by making it appear larger).

Images are commonly manipulated by controlling the distances between the object, the lens, and the screen, as well as the curvature of the lens. For example, for a given amount of curvature, when an object is closer to the lens, the focal points are farther from the lens. As a result, it is often necessary to manipulate these distances to create a focused image on a screen. Similarly, more curvature creates image points closer to the lens and a larger image when the image is in focus. This property is often described in terms of the focal distance, or distance to the focal point.

3.2.3 Electromagnetic Spectrum and Color :

Visible light is just one form of **electromagnetic radiation (EMR)**, a type of energy that is all around us. Other forms of EMR include microwaves, X-rays, and radio waves, among others. The different types of EMR fall on the electromagnetic spectrum, which is defined in terms of wavelength and frequency. The spectrum of **visible light** occupies a relatively small range of frequencies between infrared and ultraviolet light .

Whereas wavelength represents the distance between adjacent peaks of a light wave, frequency, in a simplified definition, represents the rate of oscillation. Waves with higher frequencies have shorter wavelengths and, therefore, have more oscillations per unit time than lower-frequency waves. Higher-frequency waves also contain more energy than lower-frequency waves. Higher-frequency waves also contain more energy than lower-frequency waves. This energy is delivered as elementary particles called photons. Higher-frequency waves deliver more energetic photons than lower-frequency waves. Photons with different energies interact differently with the retina. In the spectrum of visible light, each **color** corresponds to a particular frequency and wavelength (Figure 6).The lowest frequency of visible light appears as the color red, whereas the highest appears as the color violet. When the retina receives visible light of many different frequencies, we perceive this as white light.

However, white light can be separated into its component colors using refraction. If we pass white light through a prism, different colors will be refracted in different directions, creating a rainbow-like spectrum on a screen behind the prism. This separation of colors is called **dispersion**, and it occurs because, for a given material, the refractive index is different for different frequencies of light.

Certain materials can refract nonvisible forms of EMR and, in effect, transform them into visible light. Certain **fluorescent** dyes, for instance, absorb ultraviolet or blue light and then use the energy to emit photons of a different color, giving off light rather than simply vibrating. This occurs because the energy absorption causes electrons to jump to higher energy states, after which they then almost immediately fall back down to their ground states, emitting specific amounts of energy as photons. Not all of the energy is emitted in a given photon, so the emitted photons will be of lower energy and, thus, of lower frequency than the absorbed ones. Thus, a dye such as Texas red may be excited by blue light, but emit red light; or a dye such as fluorescein isothiocyanate (FITC) may absorb (invisible) high-energy ultraviolet light and emit green light (Figure 7). In some materials, the photons may be emitted following a delay after absorption; in this case, the process is called **phosphorescence**. Glow-in-the-dark plastic works by using phosphorescent material.

3.2.4 Magnification, Resolution, and Contrast :

Microscopes magnify images and use the properties of light to create useful images of small objects. • **Magnification** is defined as the ability of a lens to enlarge the image of an object when compared to the real object. For example, a magnification of 10× means that the image appears 10 times the size of the object as viewed with the naked eye.

Greater magnification typically improves our ability to see details of small objects, but magnification alone is not sufficient to make the most useful images. **.** It is often useful to enhance the **resolution** of objects: the ability to tell that two separate points or objects are separate. A low-resolution image appears fuzzy, whereas a high-resolution image appears sharp. Two factors affect resolution. The first is wavelength. Shorter wavelengths are able to resolve smaller objects; thus, an electron microscope has a much higher resolution than a light microscope, since it uses an electron beam with a very short wavelength, as opposed to the long-wavelength visible light used by a light microscope. The second factor that affects resolution is **numerical aperture**, which is a measure of a lens's ability to gather light. The higher the numerical aperture, the better the resolution.

Even when a microscope has high resolution, it can be difficult to distinguish small structures in many specimens because microorganisms are relatively transparent. It is often necessary to increase **contrast** to detect different structures in a specimen. Various types of microscopes use different features of light or electrons to increase contrast—visible differences between the parts of a specimen (see Instruments of Microscopy). Additionally, dyes that bind to some structures but not others can be used to improve the contrast between images of relatively transparent objects (see Staining Microscopic Specimens).

- Light waves interacting with materials may be **reflected**, **absorbed**, or **transmitted**, depending on the properties of the material.
- Light waves can interact with each other (**interference**) or be distorted by interactions with small objects or openings (**diffraction**).
- **Refraction** occurs when light waves change speed and direction as they pass from one medium to another. Differences in the **refraction indices** of two materials determine the magnitude of directional changes when light passes from one to the other.
- A **lens** is a medium with a curved surface that refracts and focuses light to produce an image.
- Visible light is part of the **electromagnetic spectrum**; light waves of different frequencies and wavelengths are distinguished as colors by the human eye.
- A prism can separate the colors of white light (**dispersion**) because different frequencies of light have different refractive indices for a given material.
- **Fluorescent dyes** and **phosphorescent** materials can effectively transform nonvisible electromagnetic radiation into visible light.

3.3 Absorption of light :

Light absorption is the process in which light is absorbed and converted into energy. When electrons absorb energy, they become 'excited' and move to higher energy levels which are further away from the nucleus. Electrons don't like being in an excited state, and so fall back to their original energy level very quickly – they release a packet of energy called a photon when they do this. Absorption and emission spectroscopy are two methods used to study (Fig.18 a).

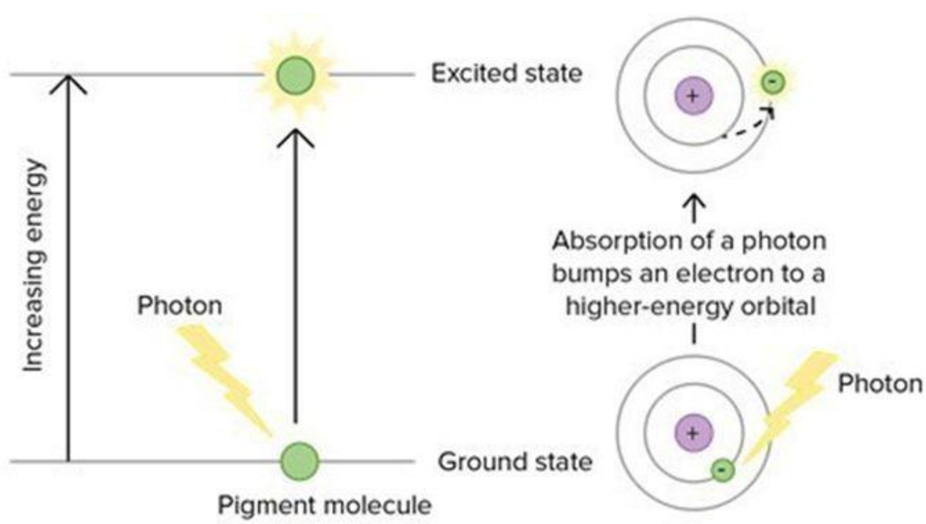


Fig 18 a Absorption of Photon

Electrons can only exist in **discrete** energy levels (these can also be called electron shells) – they can't exist halfway between. The lowest energy level that an electron can be in is called the **ground state**. For an electron to move from a lower energy level to a higher energy level, it must absorb a set amount of energy because energy levels are **quantised**. This means that the energy absorbed by the electron must be exactly the same as the energy difference between the

two levels. When an electron absorbs energy, is it promoted to a higher energy level further away from the nucleus of the atom and is described as being ‘**excited**’ (Fig.18 b)

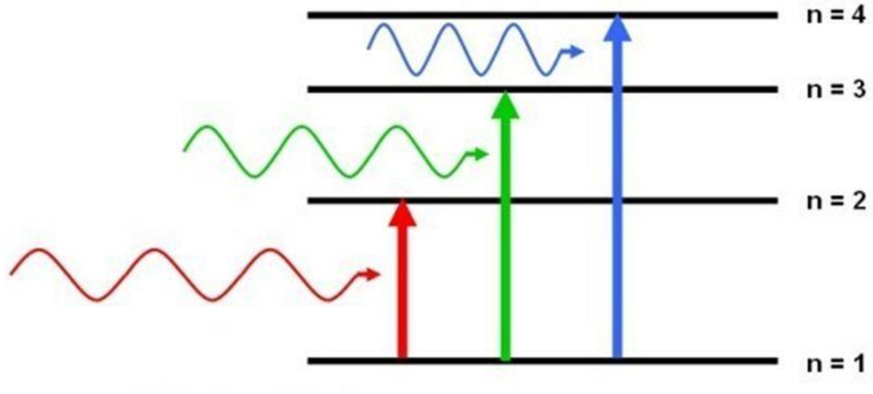


Fig 18 b Excited state of Electrons

Electrons don’t like being in an excited state. This means that after becoming excited and moving to a higher energy level, they soon fall back to their original energy level. However, to do this, they have to release a packet of energy – this is called a photon. The size of the photon released is exactly equal to the size of the jump the electron had to make in the first place.

Absorption Spectroscopy

Absorption spectroscopy is a technique used to measure the absorption of energy. The absorption spectrum of a certain material is shown by a continuous band of color with black lines between them. The coloured parts represent the total light that is focused on the material. The black lines show an absence of this light – these are the parts of the spectrum where the electrons have absorbed the light photons.

3.3.1 Absorption Spectroscopy:

Absorption spectroscopy is a technique used to measure the absorption of energy. The absorption spectrum of a certain material is shown by a continuous band of color with black lines between them. The colored parts represent the total light that is focused on the material. The black lines show an absence of this light – these are the parts of the spectrum where the electrons have absorbed the light photons.(Fig.19 a)



Fig 19 a : Absorption Spectrum

There are two types of absorption spectroscopy: atomic and molecular. Atomic absorption spectroscopy is the method of producing a spectrum when free atoms absorb different wavelengths of light – this is usually used for gases. Molecular absorption spectroscopy is the method of producing a spectrum when whole molecules absorb different wavelengths of light (usually ultraviolet or visible). Absorption spectrums are the exact opposite of emission spectrums.

3.3.2 Emission Spectroscopy:

Emission spectroscopy is used to measure the photons released when an electron falls to a lower energy level *after* becoming excited. The emission spectrum of a certain material is shown by a black band with separated coloured lines. These colours lines are the parts of the spectrum where photons have been released from the electrons when they fall to a lower energy level.(Fig.19 b)



Fig 19 b : Emission Spectrum

There are two types of emission spectroscopy: **line** and **continuous**. When the spectrum is shown as lots of lines separated by black spaces, it is a line emission spectrum. When the spectrum is shown as lots of colours in one particular wavelength, it is a continuous emission spectrum. Emission spectroscopy is used to identify a substance because the energy released when the electrons fall back to their ground state is different for every substance.
Emission spectrums are the **exact opposite of absorption** spectrums.

3.4 Energy states of atoms :

3.4.1 Atomic Energy Levels :

Atoms constitute the building blocks of all materials in existence. In these atoms, there is a central portion called nucleus (N in Figure 18) which consists of protons and neutrons, around which revolves the particles called electrons. Next, it is to be noted that all the electrons constituting the considered material do not revolve along the same path. However this even does not mean that their revolutionary paths can be random. That is, each electron of a particular atom has its own dedicated path, called orbit, along which it circles around the central nucleus. It is these orbits which are referred to as energy levels of an atom.

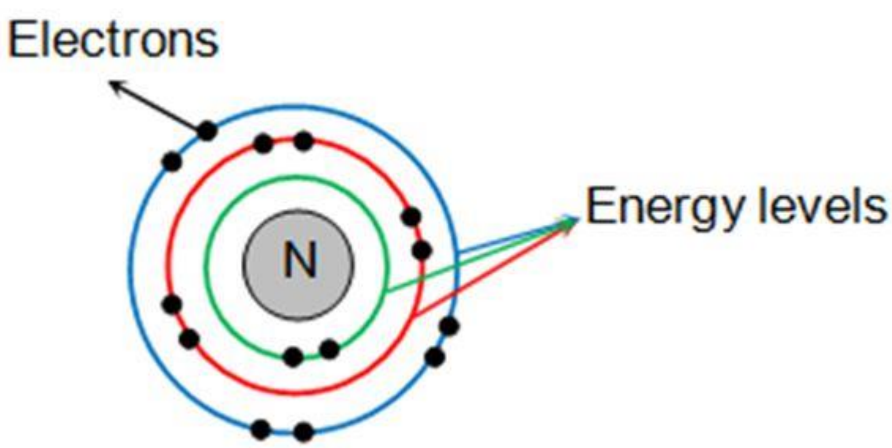


Fig 20a : Internal Structure of a Typical Atom

This is because, each of them possess a dedicated amount of energy which is expressed in

terms of an integral multiple of the equation $E = h\upsilon$
Where h is the Planck's constant and υ is the frequency.

Figure 20a shows the finite energy possessed by different energy states (and thereby all the electrons present in them) in electron volts (eV). From the figure, it can be seen that the energy of the electrons increases as one moves away from the center of the atom. For example, an electron in the first energy state (E_1) has an energy of -13.6 eV, that in the second (E_2) possess an energy of -3.4 eV and so on. Continuing so, one may reach a level at which the energy becomes 0 eV i.e. the energy level E_∞.(Fig.20b)

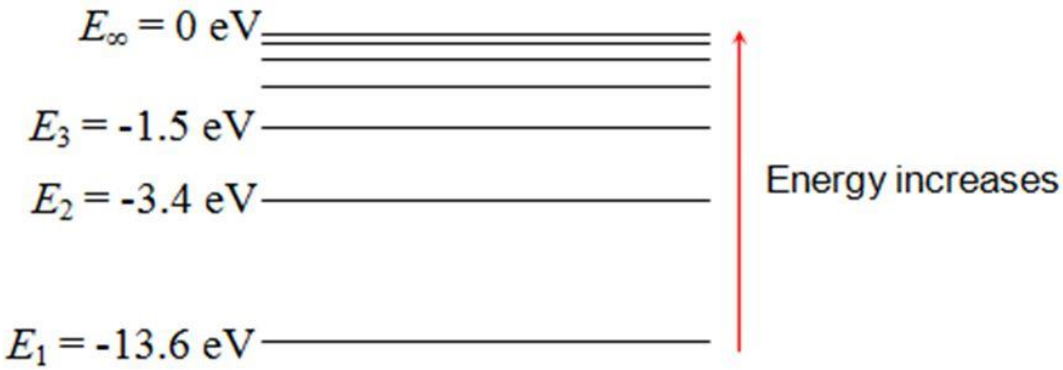


Fig 20b : Energy Levels of an Atom

Now assume that we are supplying external energy (might be in any way including that of light) to the material. This energy supplied will be absorbed by the electrons present in the atoms constituting the material. However the electrons are not let to absorb any amount of energy as they wish-for. This is because, if an electron absorbs some energy, then its net energy changes. This inturn means that the electron can no longer stay in its original energy level. Say for example, an electron in the energy state E_1 absorbs 4 eV of energy. On doing so, the net energy of the electron would increase to $-9.6\ eV (= -13.6\ eV + 4\ eV)$ due to which it can no longer stay in the energy level E_1 which has its energy as -13.6 eV. Moreover it cannot see any other level which has an energy equivalent to what it has. This makes it lose its track. On the other hand, if this electron absorbs energy of 10.2 eV, then its increased energy would be $-13.6 + 10.2\ eV = -3.4\ eV$
This is nothing but the energy possessed by the level E_2, meaning which the electron which was formerly in E_1 is now in the energy level E_2. In other words, we say that this electron has made a transition from the level E_1 to the level E_2 which inturn leads to an excited atom. However the electron cannot stay in this unstable state for long. It will soon return to its original state by making a transition from the level E_2 to the level E_1. But an important point to be noted here is the fact that while doing so, the electron emits an energy of 10.2 eV (which is same as that of the absorbed) in the form of electromagnetic waves. From the discussion presented, it is evident that the electrons are permitted to absorb (or equivalently emit) only quantized amounts of energy. The amount of this energy is nothing but the difference in the energies of the levels among which the transition occurs. Next, from Figure 2, it is seen that this difference between the energy states goes on decreasing as one moves away from E_1

i.e. $(E_4 - E_3) < (E_3 - E_2) < (E_2 - E_1)$

This means that the electrons in the outermost shells require less amount of energy to get excited than those present in the innermost shells. This is in accordance with the well known fact that the

electrons present near the nucleus are strongly bonded to the atoms rather than the ones which are present away from it.
Although we have explained the process of excitation, the same mode of argument holds good even for the case of liberation.This is because, we can assume that the electron when gets excited to the energy level with an energy of 0 eV (E_{∞}), it would be completely free from the attractive force of the atom's nucleus. It is these free electrons which contribute for conduction in the case of materials like metals.

3.5 Spin Property of Electrons:

In quantum mechanics and particle physics, **spin** is an intrinsic form of angular momentum carried by elementary particles, composite particles (hadrons), and **atomic** nuclei. **Spin** is one of two types of angular momentum in quantum mechanics, the other being orbital angular momentum.

3.5.1 Ground State :

A ground-state atom is an atom in which the total energy of the electrons can not be lowered by transferring one or more electrons to different orbitals. That is, in a ground-state atom, all electrons are in the lowest possible energy levels. (Fig.21) eg: Consider a carbon atom whose electron configuration is the following.

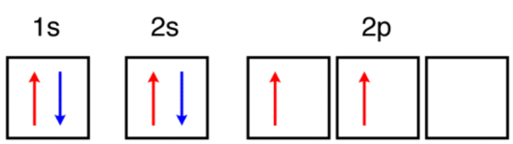


Fig.21 : Ground State of atoms

The total energy of the electrons in this carbon atom cannot be lowered by transferring one or more electrons to different orbital's. Therefore, this carbon atom is a ground-state atom.

3.5.2 Exited State of atoms and bio molecules:

An excited state is an energy level of an atom, ion, or molecule in which an electron is at a higher energy level than its ground state. An electron is normally in its ground state, the lowest energy state available. After absorbing energy, it may jump from the ground state to a higher energy level, called an excited state(Fig.22).

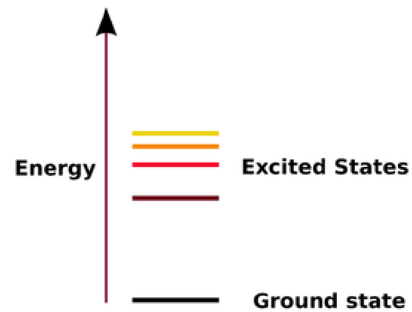


Fig.22 : Exited State of atoms

Organisms have evolved a wide variety of mechanisms to utilize and respond to light. In many cases, the biological response is mediated by structural changes that follow photon absorption. These reactions typically occur at femto- to picoseconds timescales. As the relevant time and spatial resolutions are notoriously hard to access experimentally, molecular dynamics (MD) simulations are the method of choice to study such ultrafast processes. In the simulations, a multiconfigurational quantum mechanical (QM) description (CASSCF, CASPT2) is required to model the electronic rearrangement of those parts of the system that are involved in the absorption. For the remainder, typically consisting of the apoprotein and the solvent, a simple forcefield model (MM) suffices. QM/MM gradients have to be computed on-the-fly, and surface hopping procedures are needed to model the excited state decay. In this chapter, the computational framework underlying the atomistic simulation of photochemical events is reviewed and a few representative applications are discussed that demonstrate the validity of hybrid QM/MM approaches for photo biological reactions

3.5.2.1 Effects

When an electron temporarily occupies an energy state greater than its ground state, it is in an **excited** state. An electron can become **excited** if it is given extra energy, such as if it absorbs a photon, or packet of light, or collides with a nearby **atom** or particle.

First Exited State :

If you have an atom in the ground **state** , the the **first excitation** energy is the energy separation to the lowest unoccupied orbital i.e. it is the lowest energy that can excite an electronic transfer. ... In H-atom electrons occupy shells characterized by the n . The lowest energy shell n (n=1 is the ground-**state**

Second Exited State :

The **excited state** describes an atom, ion or molecule with an electron in a higher than normal energy level than its **ground state**. The length of time a particle spends in the **excited state** before falling to a lower energy **state** varies. Short duration excitation usually results in release of a quantum of energy, in the form of a photon or phonon. let us take an example - the

degeneracy of first and second excited states. For hydrogen atom (or any other one-electron system) all orbitals from the same shell have same energy. For instance, E2s=E2p , E3s = E 3p

- The first excited state of hydrogen atom would be one in which either 2s or one of the three 2p orbitals is occupied and it will be 4-fold degenerate.
- the second excited state of hydrogen atom would be one in which either 3s or one of the three 3p or one of the five 3d orbitals is occupied and it will be 9-fold degenerate:

STUDY QUESTIONS:

Short Answer Questions

5. **State the First Law of thermodynamics.**
6. **What is Bioenergetics?**
7. **Define Redox Potential.**
8. **What is Free Energy**

Long Answer Questions

4. **Explain the Principle and Application of first two laws of thermodynamics.**
5. **Describe the coupling of chemical reactions**
6. **Write about role of NADP and NAPH in Biological Systems?**

UNIT IV

Structure

4.1. Spectroscopy :

Spectroscopy is the branch of science dealing with the study of interaction of electromagnetic radiation with matter. The most important consequence of such interaction is that energy is absorbed or emitted by matter in discrete amounts called quanta. The absorption or emission processes are known throughout the electromagnetic spectrum ranging from the gamma region (nuclear resonance absorption or the Mossbauer effect) to the radio region (nuclear magnetic resonance). The ways in which the measurements of radiation frequency (emitted or absorbed) are experimentally made and the energy levels from these are deduced comprise the practice of spectroscopy. This unit covers about the concepts of Spectroscopy with Visible and UV Spectroscopy principle and applications. The principle, instrumentation and uses of nuclear magnetic resonance (NMR) and electron spin resonance spectrometry (ESR) have been explained in detail.

4.2. Principle of Spectroscopy:

The electromagnetic spectrum for most spectroscopic purposes is considered to be consisting of the region of radiant energy ranging from wavelengths of 10 metres to I x 10-'2 centimetres. When a molecule absorbs electromagnetic radiation, it can undergo various types of excitation. This excitation may be electronic excitation, rotational excitation, excitation leading to a change in nuclear spin, excitation resulting in bond deformation and so on. Vacuum UV UV visible and near infrared ranges of spectrum are produced due to transitions which occur at the valence electron level. Far infrared range of spectrum is produced due to molecular vibration and rotation. If the energy available approaches the ionization potential of the molecule, an electron may be ejected and ionization may occur. Since each mode of excitation requires a specific quantity of energy, different absorptions appear in different regions of the electromagnetic spectrum.

The major characteristics of various spectrum regions are outlined as follows:

1. Y-ray region The 7-rays are short waves emitted by atomic nuclei involving energy changes of 10" to 10" Joules/gram atom.

2. X-ray region X-rays emitted or absorbed by movement of electrons close to the nuclei of relatively heavy atoms, involve energy changes of the order of 10000 kilo Joules.

3. Visible and ultraviolet region. This region further consists of vacuum ultraviolet, ultraviolet and visible regions.

4.3 Applications :

UV-VIS spectrophotometer is a more refined instrument and it gives a far better precision and resolution than a colorimeter. It has a wide range of applications in biological research.

1. It is used to estimate the concentration of both coloured as well as colourless solutions, which could absorb light.

2. Because of its higher sensitivity, it is used to estimate extremely small quantities of substances in a matter of a few minutes.

3. It usually does not degrade or modify the materials studied (unless a photochemical reaction occurs) and hence the materials can be recovered and reused.

4. It is also used to find out the absorption maxima of compounds with a wide range of wavelengths.

5. It offers selectivity in that each component in a solution or reaction mixture can be singled out and estimated.

6. It also enables one to follow details of fast reactions and fast enzyme kinetics.

7. It is also used to measure the growth of bacteria and yeasts and to determine the number of cells in a culture.

8. Small volumes (as small as 0.3 ml) can be used for estimation of precious samples.

Parts of a UV-VIS Spectrophotometer

Light sources A UV-VIS spectrophotometer has two light sources, a tungsten lamp for visible light, and a deuterium or a hydrogen lamp for UV lights respectively (deuterium lamp gives wider and more intense light in UV region than a hydrogen lamp). The light from the light source is composed of a wide range of wavelengths. This light is called polychromatic or heterochromatic. The polychromatic light reflected back using a plane mirror, passes through an entrance slit and through a condensing lens and falls onto a monochromatic. The monochromator disperses the light and the desired wavelength is focused on the exit slit using the wavelength selector.

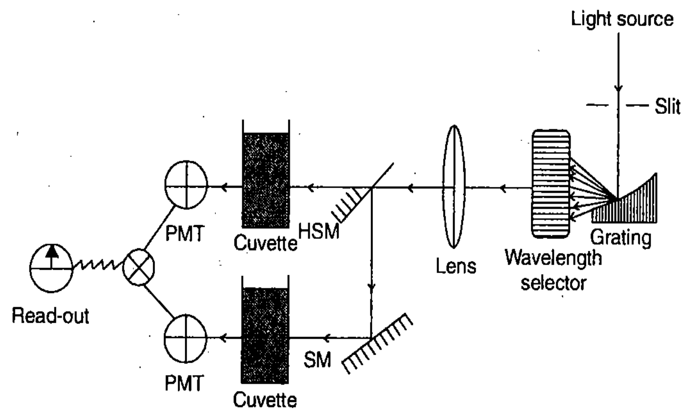


Figure 23 a. Schematic diagram of a double-beam spectrophotometer

Monochromators The monochromators which produce radiations of single wavelength are based either upon refraction by a prism or diffraction by a grating. Prisms are made of glass for visible region and of quartz or silica for UV region. A grating consists of ruled lines (as many as 2000 lines per millimetre) on a transparent or reflecting base. The resolving power of a grating is directly proportional to the closeness of these lines. Gratings are superior to prisms as they yield linear resolution of the spectrum for the entire range of wavelengths. The efficiency of monochromation is enhanced by using double monochromators in which a selected part of the spectrum from the first grating is further resolved by a second grating, resulting in a bandwidth of as low as 0.1 mm. Cuvettes.

The optically transparent cells (cuvettes) are made up of glass, plastic, silica or quartz. Glass and plastic absorb UV light below 510 nm. Hence, they cannot be used for light measurements in UV region. Silica and quartz do not absorb UV light and hence they are used for both UV and visible light measurements. Since quartz absorbs light below 190 nm, cuvettes of lithium fluoride can be used which transmit radiation down to 110 nm. Oxygen also absorbs light at wavelengths less than 200 nm. Therefore, if spectra are required in this region the apparatus must be evacuated. The standard cuvettes are made up of quartz, have an optical path of I cm, and hold a volume of 1-5 ml. Microcuvettes (0.5-0.5 ml) are used for measurement of expensive chemicals. Photocell and photomultiplier tubes A photocell (Figure 17) is a photoelectric device, which converts light energy into electrical

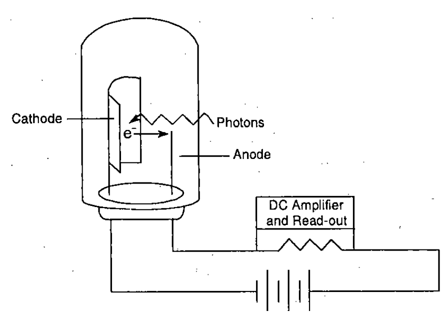


Figure 23 b. Schematic diagram of a photocell

energy, which is then amplified, detected and recorded. In photocells, the photons strike a semicylindrical photoemissive cathode in vacuum. This causes emission of electrons, which is proportional to the intensity of radiation. When a potential difference is applied across the electrodes, the emitted electrons flow to the anode wire generating a photocurrent. This current is amplified electronically and measured.

A photomultiplier tube (Figure 18), has a cathode with photoemissive surface (a selenium layer) and a wire anode. In addition to the photoemissive cathode, it also contains a circular array of nine additional cathodes called dynodes.

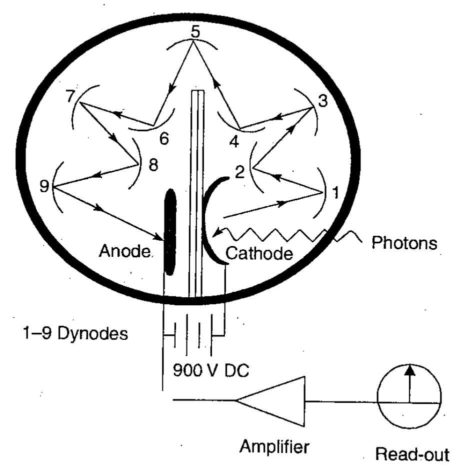


Figure 23 c :Cross-section of a photomultiplier tube

The electrons emitted from the photo emissive cathode strike dynode I, which emit several additional electrons. The electrons are accelerated towards dynode 2, which again emit several electrons. The amplified electrons flow to the anode generating a much larger photoelectric current than in a photocell.

4.4 Delayed Effect of Radiation:

The development of atomic power stations, irradiation sterilization of food, production of new chemical polymers by irradiation, application of medical instruments and use of X-ray fluoroscopy in the detection of faults in metal castings have resulted in the exposure of ionizing radiations. Furthermore increasing radioactive of them, background environment is important. Therefore, the effects accumulate and must be understood clearly after few generations. The ionizing radiations may affect the tissues, bring about mutations and hyper-sensitiveness of central nervous system. Since the effects are complex the specific effects are investigated on various aspects. From the point of view of effect the quantity in terms of rem (Roentgen equivalent man) is important. One rem is defined as the amount of damage to tissue caused by radiation of any type. This produces the same biological effects as dose 100 erg' absorbed per gram of tissue from erg of gamma radiations. One rem of damage, is produced by one rad of absorbed ergs or gamma radiations.

Biological damage is not subject to quantitative measurement. Measurment of biological damage, by nature has been done such as the LD_{50} (Lethal Dose 50). The LD_{50} is that a dose in rates which kill 50'/. of the cells or organisms irradiated' . Since irradiation change is not immediate may be said in after days or even years in the case of mammals. Generally, a limit of 50% killed within 30 days after exposure has been accepted.

Taking into the considerations of the above fact, LD 50 (30 days), for mammals is 200-1000 rads, for man it is about 400 rads. This is equivalent to 400 rems if the radiations are X or gamma radiations. The LD 50 is the useful measure for the partial body irradiation. Incidentally, it should be realized that a small amount of energy is needed to cause damage. It is the form in which this energy enters the tissue which is critical.

DNA: The nucleus is more sensitive to ionizing radiations than the cytoplasm of the cell. Among the effects of the cell DNA 'is affected much. As the order and composition of the purine and pyrimidine base pairs in a DNA molecule determine the genetic code, a change in genetic composition is caused by ionizing radiation. Pyrimidines are more sensitive than the purines. Among these, thymine is the most sensitive pyrimidine. Large doses of ionizing radiations destroy thymine as well as uracil and cytosine. Ultraviolet radiations produce thymine dimers. Since ionizing radiation causes depolymerization of DNA, it can prevent DNA replication and stop genetic transcription. This leads to mutagenesis. The high rate of mutations induced by ultraviolet light occurs at a wave length corresponding to maximum absorption of nucleic acids. U.t.-aviolet radiations are primarily absorbed by the bases. Radiations destroy DNA by various ways. Hydrogen bonds may be broken or a base may be changed in one way or other. Cross linking may appear between DNA molecule or with protein. The direct effect of the radiation is the inhibition of DNA synthesis. If it is at the time of cell division, radiation causes eel death. It is clear that DNA replication is more sensitive at the replication phase than at other periods of cell cycle. Radiation has an effect on chromosome structure. Mutation occurs in a chromosome when the DNA present fails to replicate. After a defective duplication by DNA, the mutated gene is reproduced in subsequent DNA duplication. The alteration may be in the form of an addition, deletion and inversion of bases. Structural aberrations may be produced in chromosomes by radiation at any stage of cell division; but .chromosomes at interphase are very sensitive. If a single break occurs, restoration is possible with complete recovery. Radiations can also cause breaks in chromatids. Abnormal mitosis is produced when the cell is exposed to ionizing radiation. The spindle mechanism is also affected.

One of the major effects of whole body ionizing radiation is blood cancer or leukemia. Such incidence of leukemia is reported among the survivors of Hiroshima and Nagasaki. However, dose-effect relation in leukogenic radiation is not clear. As indicated above radiation causes the increased frequency of chromosomal breaks and those individuals have a high risk of developing cancer. Ionizing radiations affect gastro intestinal tract especially epithelial cells lining the mucosa. The villi shrink due to the non-renewal of cells because of the mitotic division. As the cells are not replaced, the mucosa develops ulcer and hemorrhages In a similar way basal layer of the skin is affected by ionizing radiation. In such a case hairs are lost. Although spermatozoa are radiation-resistant, spermatogonial cells are sensitive. Small amount of radiation reduces spermatogonia and larger amounts may cause temporary or permanent sterility. Likewise, oocytes are radio sensitive. Large doses of radiation injure the brain, nervous system, heart and blood vessels.

4.5 Measurement of Radio Activity:

There are three measurement units for radioactivity: the Becquerel measures radioactivity, the Gray measures the absorbed dose and the Sievert measures the biological effects. The becquerel (Bq) is the SI derived unit of radioactivity. One becquerel is defined as the

activity of a quantity of radioactive material in which one nucleus decays per second. The activity of a source is measured in bacquerels.

This is a very small unit, and multiples are often used:

1 MBq = 1 mega Becquerel = 1,000,000 Bq

1 GBq = 1 giga Becquerel = 1,000,000,000 Bq

1 TBq = 1 tera Becquerel = 1,000,000,000,000 Bq

The radioactivity of an environment, a material or a foodstuff is given in Becquerels per kilogram or per liter.

The gray (Gy) is defined as the absorbed dose of radiation per unit mass of tissue. One gray is the absorption of one joule of radiation energy per kilogram of matter. The amount of radiation your cells absorb is measured in grays.

1 Gy = 1 joule per kilogram

Sub-multiples are often used:

1 mGy = 1 milligray = 0.001 Gy

1 μGy = 1 microgray = 0.000001 Gy

1 nGy = 1 nanogray = 0.000000001 Gy

The Sievert (Sv) is a measure of the health effects of low levels of ionizing radiation on the human body. At equal doses, the effects of radioactivity on living tissue depends on the type of radiation (alpha, beta, gamma, etc.), on the organ concerned and also on the length of exposure.

Contrary to the Becquerel, the sievert is a very large unit, and we often use sub-multiples:

1 mSv = 1 millisievert = 0.001 Sv

1 μSv = 1 microsievert = 0.000001 Sv To measure the harm done to you, we need to remember that α particles ionise very strongly, and cause 20 times more cell damage than the same dose of β particles, γ rays or X-rays. We measure the "**dose equivalent**" in **sieverts** (Sv). A dose of 1 gray of β-particles, γ-rays or X-rays will give you a dose equivalent of 1 sievert. A dose of 1 gray of α-particles will give you a dose equivalent of 20 sieverts.

Various types of instruments are used for radioactive. assay. However, the important types that are most often used in practice are mentioned here. Devices for measuring radio activity can be divided into categories according to the type of material for which they are generally used: (1) ionization chambers for gases. (2) Geiger-Muller and (3) proportional counter for liquids and solutions (4) crystal scintillation counters, for liquids and solids emitting high energy, radiations. Each device can be adopted for materials of another state of matter.

4.5.1 Geiger Muller Counter:

A typical Geiger-Muller counting system consists of a (l)Geiger {duller tube in which discharge occurs, (2) a voltage regulator that applies the electric field across that GM tube and (3) a scale or counting system for recording the discharge or counting. The useful Geiger tube (Fig. 24) consist of a cylindrical cathode and a central fine wire anode. The entire system is filled with a counting

gas that ionizes readily (e.g Q-gas 1.3% butane is helium). p-particles that originate from radioactive decay within a sample under examination must penetrate an end window to reach the electric field and its companian counting gas molecules. High energy .p-particles (p) usually penerate closely end window to result in discharge, or counts. In contrast low energy B-particle (1C, "H) have difficulty in penetrating the usual 1. 3 mg/Cm^3 mica-end windows.

Two modifications in the Geiger tube systems have been developed to overcome this problem. Firstly in the flow window systems, the mica end window is replaced with a very thin (l/1600cm) sheet of anodized plastic. The thin end window efficiently penetrates many beta-particles. Yet, because of the thinness, windows also allow slow outward diffusion of the companion counting gas from the Geiger tube. Thin window Geiger tube is therefore operated under a slight positive pressure of counting gas. The second form of modified Geiger- tube is the gas flow Geiger system. In such system the end window is eliminated and the sample is placed directly in the Geiger chamber and then the necessary.

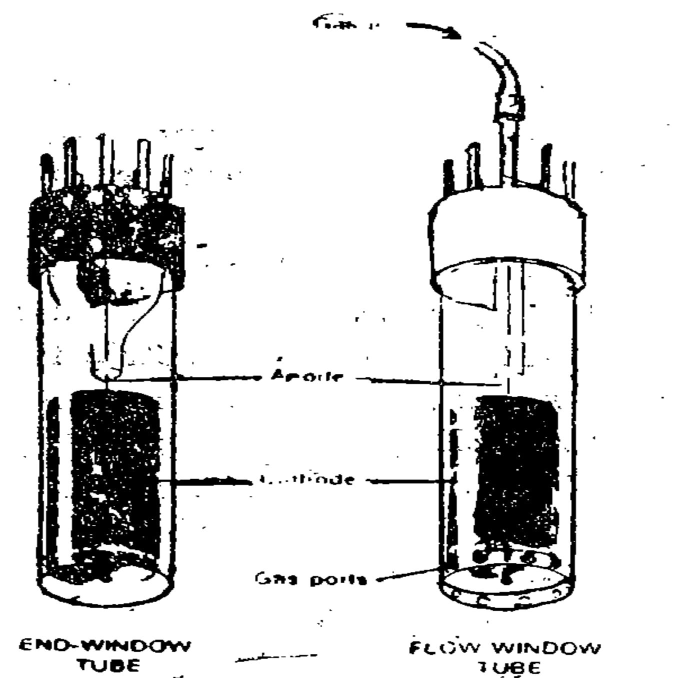


Fig . 24 :Geiger tube showing the flow window and end-window tubes.

gas supply. Therefore, gas flow Geiger system utilizes a constant flow of counting gas through Pre-flush and Geiger chambers. Several factors other than Geiger tube design influence the

counting efficiency of Geiger counters. Firstly, high rate of radiation is counted efficiently. Secondly, the position of Geiger tube near the sample tube greatly influences the Geiger counting efficiency. Thirdly, the character or thickness of the sample within the counting blanchet greatly influences the counting efficiency. Such corrections are essential for meaningful interpretation or comparison of Geiger efficiency data from different samples.

4.6 Isotopes as Tracers :

Radioactive isotopes have many useful applications in a wide variety of situations, for example, they can be used within a plant or animal to follow the movement of certain chemicals. In medicine, they have many uses, such as imaging, being used as tracers to identify abnormal bodily processes, testing of new drugs and conducting research into cures for disease.

Medical tracers : Radioactive isotopes and radioactively labeled molecules are used as tracers to identify abnormal bodily processes. This is possible because some elements tend to concentrate (in compound form) in certain parts of the body iodine in the thyroid, phosphorus in the bones and potassium in the muscles. When a patient is injected with a compound doped with a radioactive element, a special camera can take pictures of the internal workings of the organ. Analysis of these pictures by a specialist doctor allows a diagnosis to be made. The thyroid_gland, situated in the neck, produces a hormone called thyroxin, which regulates the rate of oxygen use by cells and the generation of body heat. Within each molecule of thyroxin, there are 4 iodine atoms. If a patient is made to drink a solution of sodium iodide that has been doped with radioactive iodine-131, most of it will end up in the thyroid gland. A special camera can capture the radiation emitted by the iodine-131, and an image of the gland can be constructed. An assessment can then be made about the shape, size and functioning of the gland.

Positron emission tomography (PET): A positron emission tomography (PET) scan measures important body functions, such as blood flow, oxygen use and glucose use. The information gathered helps doctors find out how well organs and tissues are functioning. Radionuclides used in PET scanning are isotopes with short half–lives, such as carbon-11 (~20 min), nitrogen-13 (~10 min), oxygen-15 (~2 min) and fluorine-18 (~110 min). These radionuclides are added into compounds normally used by the body such as glucose (or variations of glucose), water or ammonia. Such labelled compounds are known as radiotracers. In some situations, the patient is required to breath oxygen gas labelled with oxygen-15. The radionuclides used in PET decay by a process called positron emission. A positron is the antimatter version of the electron. When a positron meets an electron, an annihilation event occurs, resulting in the production of two gamma rays. The two emitted gamma rays travel in opposite directions.

4.7 Autoradiography:

The process of autoradiography basically consists of the following step's:

(i) Choice of the isotopes,
(ii) Administration of the isotopes,
(iii) Processing of the material,
(iv) Exposure of the material to emulsion,
(v) Processing of the slides and
(vi) Observations.

(i) Choice of the isotopes :

Ail matters in the universe consist of atoms. Each atom has a central nucleus around which electrically charged particles called electrons revolve. The nucleus contains two types of particles positively charged protons and charge-free neutrons. The orbiting electrons are equal to the number of protons in the nucleus. For example, the atomic structures of hydrogen, deuterium

and tritium have one electron and one proton each but the number of neutrons is variable. Most of the elements contain a combination of neutrons and protons in a stable state. But there are elements which has combinations within the nucleus that are not stable. Such elements are called isotopes. They have chemical characteristics similar to the ones which are stable elements but emit rays. They are called radioactive isotopes.

The various available isotopes are labelled as C14, P32, ¹ or I'31 They emit; any one of the three different types of particles or radiations, namely a-particles, b-particles, or g-rays (Table 24.1).a-particlesare positively charged helium nuclei that produce straight tracks inemulsion. They can be easily traced to the point of origin. In biological studies they have limited use since these a-emitters are primarily of heavy metals, b-particles are electrons having different energy levels.

(ii)Administration of the isotopes :

The isotopes are supplied in high concentrations in small ampules. The isotope is Used for either injecting into the body of.an animals or exposing it directly to the cells as in cultures and root- tips. The injection of the isotope maybe intravenou sor intraperitoneal. The intraperitoneal injection of the labelled compound is easier to carry outthan intravenous injection. It is generally carried out on an Unanaesthetized animal. All care is taken to prevent spilling or oozing o'ut of the radioactive isotope in the surroundings. The level of the precursor decreases as it gets incorporated in the tissues. After sometime, it disappears from the blood. If the diluted isotope is exposed to tissue culture cells or root-tips, the exposure is made for. a specific period.

(iii) Processing of the material :

The cells or tissues with incorporated radioactivity are processed for making cytological or histological slides. The tissues like bone marrow, kidney, testis, etc. are obtained after sacrificing the animal and the cells are air-dried or smeared,on the slides. Tissue sections are also spread on slides. The materials like testis and root- tips are used for autoradiography after mailing their squash preparations.

(iv) Exposure of the material to emulsion :

There are several methods to detect the presence of radioactivity in materials. These include scintillation counters, Geiger-Muller counter, Wilson cloud chamber and bubble chambers. But none of these can determine precisely where radioactivity is located with respect to a specific cell or an organ. It is the photographic emulsion which is used for detecting radioactivity in biological materials. The cell or organ is kept against the emulsion when the radioactivity in the form of a (or) β particles penmate the emulsion and form track marks along the path of the particle. The track is in the form of latent images. When the slides with latent images' in the emulsion is developed in a developer, silver bromide is converted into silver grains. The slides containing cells or tissues with radioactive isotope are covered with photographic emulsion. The process may be done by any one of the four methods described below:

(a) Contact method : This is the oldest and the easiest method for autoradiography. It is a crude method involving the use of a photographic film which is kept against the biological specimen. This method was used for large specimens and that too for gross localization of the isotopes.

(b) Nuclear track plate method : This method consists of using nuclear track plates supplied as glass slides having one face covered with a 10-100 mm thick emulsion. The slides are available in packages of a dozen each. These are available with several types of emulsions, viz., nuclear track alpha (NTA), nuclear track beta - I, (NTB), NTB-2, NTB-3 and nuclear track electron (NTE). All these emulsions are uniform and fine grain emulsions.

(c) Emulsion coaling method: This method consists of the use of liquid emulsion like NTA, NTB-1, NTB-3 manufactured by Kodak. Ilford also supplies emulsions under the trade names ofG-5, K-2, K- 5 and L-4. The emulsions vary in their sensitivity, initial background fog and silver grain size. For example, NTB-I has large sized grains while NTB-3 has very fine grains and is used for quantitative studies. The coating process is done in a cylindrical or flat glass tube which can accommodate not more than two slides at a time. This for economy in the use of expensive emulsion. The slides are kept in a slanting position to drain extra emulsion from the slides in dark. They are kept in light-tight boxes for exposure. The coating method is economical and practical for routine cytological studies.

(d) Stripping film method. This method was first used by Peic in .1947: It consists of coating of the sections or smears with pieces of stripping film instead ofthe liquid emulsion. Kodak AR-10 fine grain autoradiographic stripping plate or AR-50 fast autoradiographic plate or Ilford special half-tone stripping plate are used. Before sections and smears are made the slides are dipped into a 0.5% gelatin solution containing 0.05% chrome alum. This helps in a good attachment of the film to the glass slides.

STUDY QUESTIONS:

Short Questions

1. **Define Spectroscopy.**
2. **What is Delayed Effects of Radiation?**
3. **What is an Isotope?**
4. **Mention any two Application of Spectroscopy.**

Long Questions:

1. **Explain the Principle and Application of Spectroscopy.**
2. **Describe Geiger Muller Counter.**
3. **Give an account on Autoradiography.**

BLOCK II: BIOSTATISTICS

UNIT V

Structure

5.1 Definition and Scope of Biostatistics:

Biostatistics is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine, agriculture and fishery; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results. A major branch of this is medical biostatistics, which is exclusively concerned with medicine and health.

Biostatistics is the application of statistical science to research in health-related fields including medicine, biology, public health, nursing and pharmacy. The objective ofBiostatistics is to advance statistical science and its application to problems of human health and disease, with the ultimate goal of advancing statistics. The role of biostatisticians is an important one, especially when it comes to designing studies and analyzing data from research problems. Biostatisticians help in formulating the scientific questions to be answered, determine appropriate sampling techniques, coordinate data collection procedures, and conduct statistical analyses to answer those scientific questions. Biostatisticians also play vital role in the preparation of research material for publication

5.2 Collection of data:

The first step in any enquiry (investigation) is the collection of data. The data may be collected for the whole population or for a sample only. It is mostly collected on a sample basis. Collecting data is very difficult job. The enumerator or investigator is the well trained individual who collects the statistical data. The respondents are the persons from whom the information is collected.

5.2.1 Methods of Collecting Data:

- Direct personal interviews.

- Indirect Oral interviews.
- Information from correspondents.
- Mailed questionnaire method.
- Experimental Methods.

Direct Personal Interviews:

- In this method there is a face to face contact with the person from whom the information is to be obtained
- The investigator personally meets them and asks questions to gather the necessary informations.
- If person wants to collect data about the blood group of the students of a college, he would go to the college and contact the students directly and get the information.

Indirect Oral interviews:

- In this method the investigator contact third parties to get the information
- The investigator indirectly interview the persons through their friends or neighbors
- This method is preferred if the required information is on addiction or cause of fire or theft or murder etc.

Information from correspondents:

- The investigator appoints local agents or correspondents in different places to collect information
- The advantage of this method is that it is cheap and appropriate for extensive investigations
- Informations to Newspapers and some departments of Government come by this method.

Mailed questionnaire method:

- This method a list of questions is prepared and is sent to all the informants by post.
- The list of questions is technically called questionnaire.
- This method is appropriate in those cases where the informants are spread over a wide area

Experimental Method:

- In this method data are collected from the experimental results conducted in the research departments. There are two types for the collection of data they are Primary and Secondary data
-

5.4 Primary Data:

Primary data are the first hand information which is collected, compiled and published by organizations for some purpose. They are the most original data in character and have not undergone any sort of statistical treatment.
Example: Population census reports are primary data because these are collected, complied and published by the population census organization.

5.5 Secondary Data:

The secondary data are the second hand information which is already collected by an organization for some purpose and are available for the present study. Secondary data are not pure in character and have undergone some treatment at least once.

Example: An economic survey of England is secondary data because the data are collected by more than one organization like the Bureau of Statistics, Board of Revenue, banks, etc.

STUDY QUESTIONS:

Short Questions:

1. **Define Bio Statistics**
2. **Mention briefly the Scope of Biostatistics**
3. **What is a Data?**
4. **Write any two methods of collection of data?**

Long Questions

1. **Give an account on collection of data.**
2. **Explain Primary and Secondary data.**

UNIT VI

Structure

6.1 Types of Sampling:

- **Random sample** gives every member of the population an equal chance of being selected. No one in the population is Favoured over other in the selection process
- **Biased sample or Nonrandom sample** does not provide equal opportunity for all members of the population of being selected. Sample is drawn with a purpose.
- **Stratified random sampling – Polygon**

6.1.1 Random Sampling:

Random sampling is important because it allows us to apply the laws of probability to sample data, and to draw inferences about the corresponding populations. B. Weaver (31-Oct-2005) Probability & Hypothesis Testing 2 Sampling With Replacement A sample is random if each member of the population is equally likely to be selected each time a selection is made. When N is small, the distinction between with and without replacement is very important. If one samples with replacement, the probability of a particular element being selected is constant from trial to trial (e.g., 1/10 if N = 10). But if one draws without replacement, the probability of being selected goes up substantially as more subjects/elements are drawn. e.g., if N = 10 Trial 1: p(being selected) = 1/10 = .1 Trial 2: p(being selected) = 1/9 = .11111 Trial 3: p(being selected) = 1/8 = .125 etc. Sampling Without Replacement When the population N is very large, the distinction between with and without replacement is less important. Although the probability of a particular subject being selected does go up as more subjects are selected (without replacement), the rise in probability is minuscule when N is large. For all practical purposes then, each member of the population is equally likely to be selected on any trial. e.g., if N = 1,000,000 Trial 1: p(being selected) = 1 / 1,000,000 Trial 2: p(being selected) = 1 / 999,999 Trial 3: p(being selected) = 1 / 999,998 etc.

6.1.2 Stratified Sampling:

In a stratified sample the sampling frame is divided into non-overlapping groups or strata, e.g. geographical areas, age-groups, genders. A sample is taken from each stratum, and when this sample is a simple random sample it is referred to as stratified random sampling.

Advantages:

- Stratification will always achieve greater precision provided that the strata have been chosen so that members of the same stratum are as similar as possible in respect of the characteristic of interest. The bigger the differences between the strata, the greater the gain in precision. For example, if you were interested in Internet usage you might stratify by age, whereas if you were interested in smoking you might stratify by gender or social class.
- It is often administratively convenient to stratify a sample. Interviewers can be specifically trained to deal with a particular age-group or ethnic group, or employees in a particular industry. The results from each stratum may be of intrinsic interest and can be analysed separately.
- It ensures better coverage of the population than simple random sampling.

Disadvantages:

- Difficulty in identifying appropriate strata.
- More complex to organise and analyse results.

Choice of Sample Size for each Stratum:

In general the size of the sample in each stratum is taken in proportion to the size of the stratum. This is called **proportional allocation**. Suppose that in a company there are the following staff:

male, full time	90
male, part time	18
female, full time	9
female, part time	63

and we are asked to take a sample of 40 staff, stratified according to the above categories.

The first step is to find the total number of staff (180) and calculate the percentage in each group.

% male, full time = (90 / 180) x 100 = 0.5 x 100 = **50**
% male, part time = (18 / 180) x100 = 0.1 x 100 = **10**
% female, full time = (9 / 180) x 100 = 0.05 x 100 = **5**
% female, part time = (63/180)x100 = 0.35 x 100 = **35**

This tells us that of our sample of 40,

50% should be male, full time.
10% should be male, part time.
5% should be female, full time.
35% should be female, part time. **50%** of 40 is 20.
10% of 40 is 4.
5% of 40 is 2.
35% of 40 is 14.

Sometimes there is greater variability in some strata compared with others. In this case, a larger sample should be drawn from those strata with greater variability.

6.2 Variables

6.2.1 Types of Variables :

Variable is a quantity which can vary from one individual to another. For example, animals of same species may differ in their length, weight, age, sex, etc. These characteristics are variables. Thus, variable can be defined as "the characteristics by which individuals differ among themselves. The particular values of a variable are termed as variate or values. Variables may be of two types (Fig. 2).

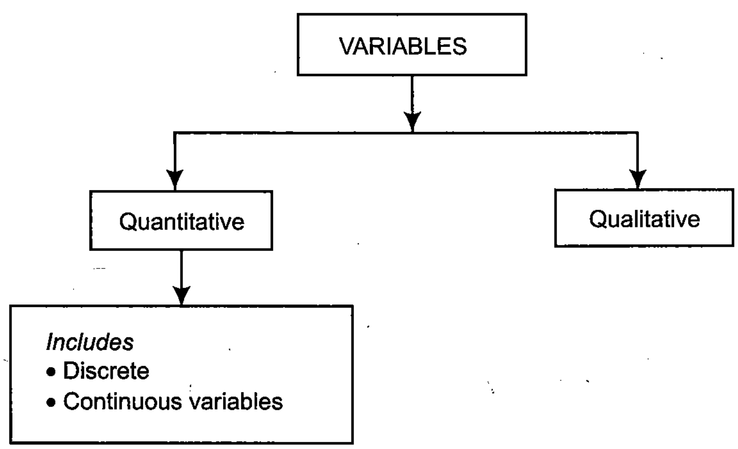


Fig. 25 Types of variables.

6.2.1.1 Quantitative variable

It is a characteristic which can be measured on a scale in some appropriate units, e.g., measurement of age, weight, length, etc. Quantitative variables can be further sub- divided into two types:

(i) Discontinuous or discrete Variable is one which is incapable of taking all possible values, e.g., the number of rooms in a house or the number of persons in the family can take only the integral values such as 2, 3, 4, etc. Here a count of 2 is not possible.

(ii) Continuous variable is one which can take any numerical value within a certain range, e.g., the height of a child at various ages when he grows from 120 cm tol50 cm, assumes all possible' values within the limit even in fractions.

6.2.1.2 Qualitative variable :

It is one which is immeasurable and inexpressible in magnitude. It can be expressed in qualities which are called attributes, e.g., color of flowers, texture of leaves, etc.

STUDY QUESTIONS:

Short Questions

1. **Define Random Sampling**
2. **Define Discontinuous Variable**
3. **What is Stratified Random Sampling?**
4. **What is Qualitative Variable?**

Long Questions

1. **Explain the concept of Sampling.**
2. **Describe the Types of Variables with suitable example.**

UNIT VII

Structure

7.1 Presentation of data

- 7.1.1 Bar diagram
- 7.1.2 Histogram
- 7.1.3 Polygon
- 7.1.4 Pie diagram

7.1 Presentation of the data:

A large no. of diagrammatic devices are used to present statistical data. Examples: One dimensional – Line & Bar diagram. Two dimensional – Circles & Pie diagram.

7.1.1 Bar diagram:

Bar diagrams are commonly used to present the data as illustrations. They consists of a group of equidistant rectangles, one for each group or category of the data in which the values represented by the length are height of the rectangles. The bar diagrams include simple bar diagram, component bar diagram, percentage bar diagram, Multiple bar diagram and bilateral bar diagram.

Simple bar diagram. It is the simplest and more frequently used diagram for the comparison of two or more items or values of a single variable.

For example, the data relating to birth, death, growth, yield, etc., for different periods may be presented by bar diagrams.

Model problem (Fig.1):

Months	S	O	N	D
No.ofedaphic	150	300	500	400

Arthropod

(No X $10^2/M^2$)

Component bar diagram

The component bar diagrams are used of the total magnitude if the given variable is to be divided into various parts or components.(Fig.26)

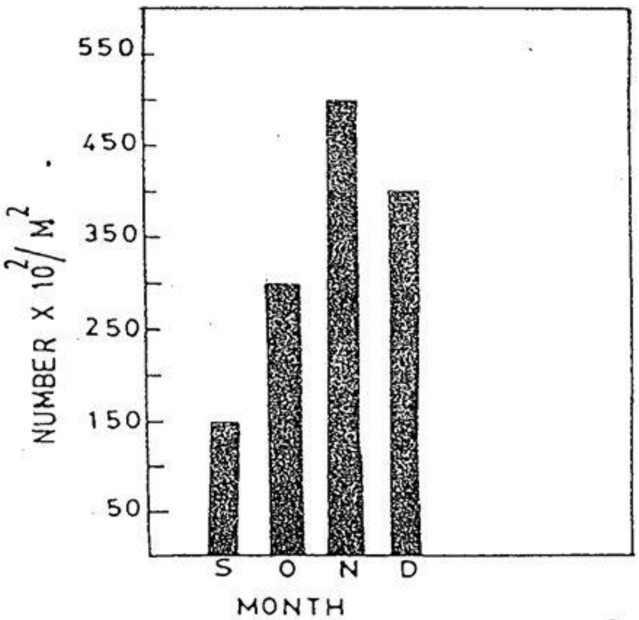


Fig.26 :Bar Diagram

7.1.2. Histogram:

One of the most important and useful methods of presenting frequency distribution of continuous series is known as histogram. In this, the magnitude of the class interval is plotted along X-axis and the frequency on Y-axis. Each class has lower and upper values. This gives us two equal lines representing the frequencies of upper ends of the lines are joined together. This process will give us rectangles, as there are classes and the heights of the rectangles are proportional to their frequencies. When class-intervals are un-equal, a correction for un-equal class-intervals must be made. Example (Fig.27).

Draw a histogram for the following data :

Length (cm)	4-5,	5-6,	6-7,	7-8,	8-9	&	9-10
No. of shells.	5	10	16	18	12		6

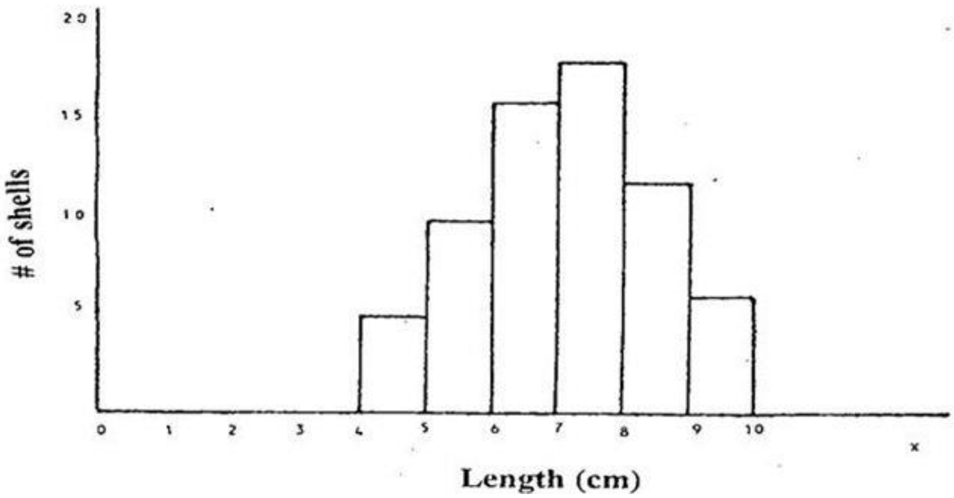


Fig.27 :Histogram

7.1.3 Polygon:

A frequency polygon is a graphical form of representation of data. It is used to depict the shape of the data and to depict trends. It is usually drawn with the help of a histogram but can be drawn without it as well. A histogram is a series of rectangular bars with no space between them and is used to represent frequency distributions.(Fig.28)

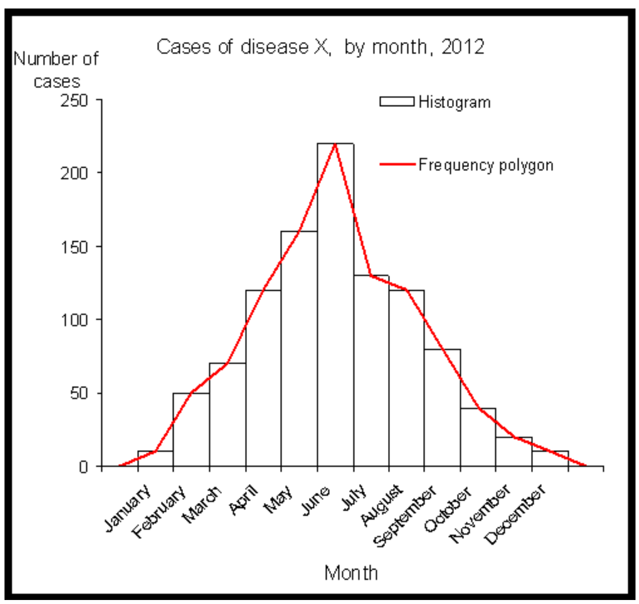


Fig.28 : Polygon

7.1.4 Pie Diagram:

A circle may be divided into various sections of segments representing certain proportions or percentage to the total is known as an angular or pie diagram.

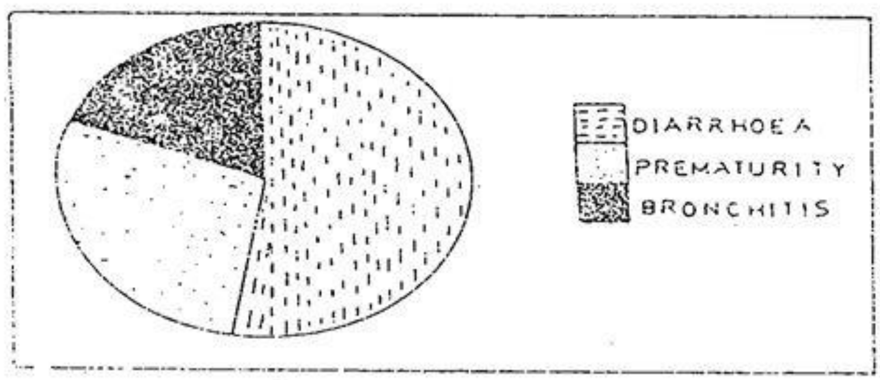


Fig.29 : Pie Diagram

Steps in construction of pie diagram

i. Convert each of the component value into percentage of the representative

total.

ii. A circle has 360° out of which 1% is equal to 360/100 = 3.6, using this, value,

the percentage of the components parts obtained in (i) can be converted to degrees by multiplying each of them by 3.6.

iii. If only one category is to be studied, the circle may be drawn on any radius.

If two or more sets of data are to be represented simultaneously for comparative studies, then the radii of the corresponding circles are to be proportional to the square root of their total magnitudes.

iv. If a protractor is divided into 100 equal parts instead of 360°, then the angle

represting any desired percent can be read directly.

v. Different sectors representing various component parts should be distinguished from one another by using different shades, dottings, etc.(Fig.29).

Causes of death	Members	Angle of the circle
Diarrhoea & enteritis	60	$\frac{60}{320}$ X 360 = 68°
Prematurity & atrophy	170	$\frac{170}{320}$ X 360 = 191°
Bronchitis & pneumonia	90	$\frac{90}{320}$ X 360 = 101°
	Total	360°

STUDY QUESTIONS:

Short Questions

1. **What is Presentation of data?**
2. **What is Bar Diagram?**
3. **Define Histogram**
4. **What is a Polygon**

Long Questions

1. **Explain the various types of Line and Bar diagram.**
2. **Explain Pie Diagram with a suitable example.**

BLOCK III: MEASURES OF CENTRAL TENDENCY AND MEASURE OF DISPERSION

UNIT VIII

Structure

8.1 Mean:

8.1.1 Arithmetic Mean (a.m.):

The common average of many individual values of observations or items is referred to as t arithmetic mean. It is the number obtained by dividing the sum of values of all the items in a sen bv the total number of items in that series.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots X_n}{n} = \frac{\Sigma X}{n}$$

In this formula:
X (read as 'X-bar') denotes arithmetic mean and

X,, Xg, X3, ...etc. are different values of variable X.

n is the number of observations of variable X.

Symbol S is Greek letter sigma. It denotes sum /.e.SX is the sum of all values of X.

Types of Arithmetic Mean

1. Simple Arithmetic Mean: In calculating simple average, all items of a series are given equal importance.

2. Weighted Arithmetic Mean: In weighted arithmetic mean, the average reflects the relative importance of different items of a series.

Calculation of Simple Arithmetic Mean

Different formulae are used for calculating arithmetic mean of ungrouped or raw data and of grouped data.

Arithmatic Mean of Ungrouped or Raw Data

We know that ungrouped data consists of individual observations. This type of arithmetic mean on the average is calculated by summing up all the individual observations or measurements of a sample and dividing the total by the number of items, observations or measurements.

A. Direct Method**:** As discussed above, the simple Arithmetic Mean can be calculated by using the following formula:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots X_n}{n} = \frac{\Sigma X}{n}$$

$$\text{Arithmatic mean} = \frac{\text{Sum of observations}}{\text{No. of observations}}$$

1. The individual observations are represented by
$x_1, x_2, x_3, \ldots\ldots\ldots x_n$

2. The sum of individual observations is represented by SX

3. The number of observations by n and

4. Their mean by X (say 'X bar')

$$\text{Then Mean } \bar{X} = * \frac{X_1 + X_2 + X_3 + \ldots X_n}{n} = \frac{1}{n}\sum_{x_i}^{n} = \frac{\Sigma X}{n}$$

Here, symbol S means summarization.

Example 1: In a respirometer, the oxygen concentration in four cases was recorded to be:

1. 14.9% 2. 10.8% 3. 12.3% 4. 23.3%.

Calculate the arithmetic mean of the above values.

Solution:

Step 1. The sum total of these observations = 61.3%

Step 2. The mean oxygen percentage calculated in these four observations will be:

63.3/4 = 15.325 Ans.

Example 2: Find the arithmetic mean of the marks obtained by 10 students of a class in mathematics in a certain examination. The marks obtained are:

25, 30, 21, 55, 47, 10, 15, 17, 45, 35.

Solution:

Let X be the average marks obtained.

Sum of all the observations ΣX = 25 + 30 + 21 + 55 + 47 + 10 + 15 + 17 + 45 + 35 = 300
Number of students n = 10

Arithmetic mean $\times = \frac{\Sigma X}{n} = \frac{300}{10} = 30 Ans$

Example 3: Calculate the arithmetic mean of 10 observations related to the lengths (in cm of radishes:

11, 12, 10, 10.5, 9.8, 10.7, 11.2, 9.3, 12, and 8.9.

Solution:

Step: Sum of the observations

$\Sigma X = 11 + 12 + 10 + 10.5 + 9.8 + 10.7 + 11.2 + 9.3 + 12 + 8.9 = 105.4$

Step: Number of observations (n) = 10

Step: Arithmetic mean $\times = \frac{\Sigma X}{n} = 105.4/10 = 10.54 Ans$

B. Short-cut Method : Short cut method for calculating arithmetic mean is used when the number of itmes the series is very large. The formula used is as follows:

Formula = $\times = A + \frac{\Sigma d}{n}$

Here X is the actual arithmetic mean

A is the assumed arithmetic mean

d is the deviation of items from the assumed mean, /.e. d = (X-A

Σd is the sum of deviations from the assumed mean

n is the total number of observations.

Steps in Calculation:

Step 1. Assumed mean (A) of the series is calculated by dividing the total of maximum and the minimum values of the items of a series with two.

Step 2. Deviation(d) of different values from the assumed mean is calculated by subtracting the assumed mean from the actual value (i.e. X_1, A, X_2-A....). Step3. Sum of all these deviations (Σd) is calculated by addition.

Step 4. All these values are placed in the above formula: $\times = A + \frac{\Sigma d}{n}$

Example 4**:** The table given below shows the number of colonies of bacteria grown on ten agar plates Calculate the arithmetic mean by using short cut method.

Plate no.1.	1	2	3	4	5	6	7	8	9	10
No. of Colonies	60	70	80	95	100	110	115	130	140	160

Solution:

Calculation of arithmetic mean using short-cut method:

Serial No of Plates	No. of Bacterial colonies per plate	Assumed Mean	Deviation from Assumed Mean
1	60		60- 110 =-50
2	70		70- 110 =-40
3	80		80- 110 =-30
4	95		95- 110 =-15
5	100	$\frac{60+160}{2}=\frac{220}{2}$	100- 110 =-10
6	110	=100	110- 110 = 0
7	115		115- 110 = 5
8	130		130- 110 = 20
9	140		140- 110 = 30
10	160		160- 110 = 50
n =10	ΣX=1060		Σd = - 145-(+105)-40

Step 1. Assumed Mean A = $\frac{60+160}{2}=\frac{220}{2}=100$

Step 2. Σd = - 40 (calculated as shown in the table)

Step 3. n = 10

Step 4. Putting the values in the formula = $\times = A+\frac{\Sigma d}{n}$

$$110+\frac{-40}{10}=100-\frac{40}{10}=106 Ans.$$

Arithmetic Mean of Grouped Data

It can be calculated by the use of following formula:

$$Formula: X = \frac{f_1X_1 + f_2X_2 + f_3X_3 + f_4X_4 .... f_nX_n}{f_1 + f_2 + f_2 + f_4 + ... f_n}$$

$$or X = \frac{\Sigma fx}{\Sigma f}$$

f' = frequecny

X = value of each item

1. Arithmetic Mean of Grouped Data (Discrete Series): In the case of discrete series, arithmetic mean of group data is calculated by the use of following formula:

$$\times = \frac{\Sigma fx}{\Sigma f} or \frac{1}{N} \Sigma fx$$

Here N = Σf. It is sum of all the frequencies.

Steps in Calculation**:**

When a particular value occurs more than once, the arithmetic mean of group data is calculated:

1. By multiplying each value (X) with the corresponding frequency of its occurrence (obtaining of different observations: f_1X_1, f_2X_2…,f_n X_n).

2. Adding all the multiplication products to obtain (Σfx).

3. Divide this total value £ fx by total number of observations or total frequencies:

$$\frac{\Sigma fx}{\Sigma f} or \frac{\Sigma fx}{N}$$

Example 5: Find the mean from the following data:

Marks(X) :	5	10	15	20	25	30	35	40
No. of Students (f):	5	7	9	10	8	6	3	2

Mark obtained by different number of student

Marks(X)	**No. of Students (f)**	fx
5	5	25
10	7	70
15	9	135
20	10	200
25	8	200
30	6	180
35	3	105
40	2	80
	$\Sigma f = 50$	$\Sigma fx = 995$

$$\bar{X} = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{995}{50} = 19.9 \quad \textbf{Ans.}$$

2. Arithmetic Mean of Group Data (Continuous Series): When there is a continuous class distribution from X_0-X_1, X_1-X_2, X_2-X_3,...with their corresponding frequencies as f_1,f_2,f_3..., the arithmetic mean is calculated by using the following formula:

Formula $$\times = \frac{\Sigma fm}{\Sigma f}$$

Here, m = the mid value of various classes

f = the frequency of each class

Σf or Sf = the total frequency

Steps in Calculation**:**

1. Firstly, the mid value of each class is calculated.

2. Multiply each mid value by respective frequency (f) to obtain value of fm.

3. Add all these fin values to obtain Σfm or Σfx.

4. Divide this value or Σfm or Σfx by the sum total of all frequencies, i.e., total number of observations.

Example 6 Values of fecundity (rate of reproduction) of 50 fishes of a species of fish are given in a frequency table. Calculate the mean value of fecundity.

Table 2 Values of fecundity of 50 fishes of a species

$$\textbf{Calculation :} \quad \bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{1875}{50} = 37.5 \quad \textbf{Ans.}$$

8.1.1.1 Merits of Arithmetic Mean:

1. Certainity: Arithmetic mean is rigidly defined. So its value is always definite and certain. Mean can never be biased.

2. Simplicity: Arithmetic mean is easy to calculate and simple to understand.

3. Stability: Arithmetic mean is a relatively stable measure. It is least affected by fluctuations of sampling.

4. Based on Observations: Arithmetic mean is based on all the observations of a series. Therefore, it is' a most representative measure.

5. Algebraic Treatment: Arithmetic mean is capable of further algebraic treatment. Because of this attribute arithmetic mean is extensively used in statistical analysis.

6. Basis of Comparison : Arithmetic mean is the best measure for comparing two or more series of data.

7. Balance : Arithmetic mean balances the value on either side.

8.1.1.2 Demerits of Arithmetic Mean:

1. Affect of Extreme Values : Since arithmetic mean is the average of all the values of a series, it is greatly affected by extreme fluctuations. Thus, it is not a true representative value of all the items of the series.

2. Problem in Case of Incomplete Data : Arithmetic mean can not be calculated unless all the items of the series are known.

3. Mean Value may not Figure in the Series : Arithmetic mean value sometimes does not appear in the series. For example, the arithmetic mean of 4, 8, 15 and 21 is 12 but it is not present in the series.

4. Misleading Conclusions: Arithmetic mean sometimes provides misleading conclusions. For example, 3 workers in factory A are getting daily wages of Rs 60, 70 and 80 respectively, while in factory B, three workers get daily wages of Rs 60,30 and 120. The average daily wages of workers in both the factories comes to Rs 70. These results are misleading, because the workers in two factories have same average of income but different income structure.

5. Absured Results : Arithmetic average sometimes gives results which are absurd and unacceptable. For example, in five families, the number of children is found to be 2,3,4,3 and 5. The average number of children per family comes out to be:

$$\frac{2+3+4+3+5}{5} = \frac{17}{5} \text{ 3.4 children}$$

This result is absurd because children cannot be divided into fractions.

6. Arithmetic mean cannot be used for small number of classes.

7. Arithmetic mean cannot be used for qualitative characterstics such as colour of flowers, sweetness of orange or darkness of the color.

8.1.2 Combined Arithmetic Mean:

When two or more distributions are given with their number of items and their arithmetic means, their combined mean is called combined arithmetic mean. If we are given the mean of two series and their size, then the combined mean for the resultant series can be calculated by the formula :

$$\bar{\bar{X}} = \frac{\bar{X}_1 N_1 + \bar{X}_2 N_2}{N_1 + N_2}$$

(Read X double bar)

X = Combined arithmetic mean
X_1 = Arithmetic mean of first distribution

N_1 = No. of items of first distribution

X_2= Arithmetic mean of 2nd distribution

N_2= No. of items in 2nd distribution.

Example 1**:** The mean age of 40 students is 16 years and the mean age of another around of 60 students is 20 years. Find out the mean age of all the 100 students combined together.

Solution : Let the first distribution be represented by affix 1 and the second distribution is represented by affix 2.

$X_1 = 16$ $N_1 = 40$

$X_2 = 20$ $N_2 = 60$

$$\bar{\bar{X}} = \frac{\bar{X}_1 N_1 + \bar{X}_2 N_2}{N_1 + N_2}$$

$$= \frac{(16 \times 40) + (20 \times 60)}{40 + 60}$$

$$= \frac{640 + 1200}{100} = \frac{1840}{100} = 18.4$$

The combined mean of age of 100 students = 18.4 Ans.

1.Calculation of Mean by Assumed Mean Method

Mean can be calculated by using another method in which an arbitrary reference point is taken. This arbitrary reference point istermed-as Assumed Mean or Provisional Mean. This is the short-cut method which is applied when the frequencies and the values of variables are very large and it becomes very difficult to compute the arithmetic mean. The arithmetic mean is given by the formula:

(a) In Case of Ungrouped Data ;

$$\times = M + \frac{\Sigma d}{N}$$

where, X = Mean of observations

M' = Assumed mean

d = Deviation from the arbitrary reference point

N = Σ*f*= Total number of observations

Example 2**:** Find the mean by short-cut method using the following data:

59, 65, 71, 67, 61, 63, 69, 73

Solution: Let us take 65 as assumed mean and prepare the following table:

X	**d =X-65**
59	59 – 65 = 6
65	65 – 65 = 0
71	71 – 65 = +6
67	67 – 65 = +2
61	61 – 65 = - 4
63	63 – 65 = -2
69	69 – 65 = +4
73	73 – 65 = +8
N =8	**Σd =8**

$$X = M + \frac{\Sigma d}{N} = 65 + \frac{8}{8} = 66 \text{ cm Ans.}$$

(b) In Case of Grouped Data :

$$\times = M + \frac{\Sigma fd}{N}$$

Where, *f* = Frequency

*f*d = product of the frequency and the corresponding deviation.

(c) Data with Equal Class Intervals**:**

In case of frequency table with equal class intervals, the arithmetic mean is given by the formula: $\times = M + \frac{\Sigma d}{N} i$

Where i = with of the Class Intervals:

Example 3**:** Find the mean from the following data using assumed mean method :

Marks :	0-10	10-20	20-30	30-40	40-50
No. of students :	42	44	58	35	26

Solution**:**

Put the values in the following formula:

$$\times = M + \frac{\Sigma fd}{N} i = 35 + \left[\frac{-216}{220} X10\right]$$

=35-9.8 = 25.2 Ans.

8.1.3 Geometric mean (G.M.)

The geometric mean of a set of data for n observations is the xth root of their product. If, X_1, X_2, X_3 ....X_n are the given n observations then geometric mean is

$$G.M. = n\sqrt{X_1.X_2.X_3...X_n}$$

$$=(X_1, X_2, X_3, \ldots X_n)^{1/n}$$

If n = 2, i. e., the number of observations are two only, the GM can be computed by taking the square root of their product. Only the geometric mean can be computed by taking the square root of their product.

Say, G.M. = $\sqrt{4}$ = 1.41 and of $\sqrt{64}$ = 8

But if n = more than 2, then the computation of the n^{th} root is difficult. In that case the calculations are made by making use of the logarithms :

$$\log G.M. = \frac{1}{n} \log (x_1, x_2, x_3, \ldots\ldots\ldots x_n)$$

$$= \frac{1}{n} (\log X_1 + \log X_2 + \log X_3 \ldots\ldots\ldots \log X_n)$$

$$= \frac{1}{n} \Sigma f \log$$

Taking antilog of both sides, we finally obtain :

$$G.M. = \text{Antilog } \frac{1}{n} \log X$$

In case of frequency distribution:

$$\log G.M. = \frac{1}{n} \Sigma f \log$$

8.1.3.1 Merits of Geometric Mean:

1. It is based on all observations.

2. Arithmetic mean has a bias for higher values whereas Geometric Mean has bias for smaller observations.

3. It is not affected much by fluctuations of sampling.
4. It is useful in averaging ratios, percentage rate of increase and decrease between two persons.

8.1.3.2 Demerits of Geometric Mean :

1. Geometric Mean is a mathematical character. It is not easy to understand or to calculate for non-mathematical persons.

2. If any observation (X_1, X_2), is zero. Geometric Mean would be zero and if any one of the observations is negative, the Geometric Mean becomes imaginary.

Example **:** Find the average rate of increase in tiger population which in the first decade had increased by 20%, in the second decade by 30% and in the third by 40%.

Solution: Here we have to determine the rate of increase in population. The appropriate average to be computed is G.M. and not the arithmetic mean. Hence, the average percentage rate of increase in the tiger population per decade over the entire period = 129.7-100 = 29.7.

$$G.M. = \text{Antilog } \frac{1}{n} \log X$$

$$=\text{Antilog } \frac{6.3392}{3} = (2.1131) = 129.7 \text{ Ans}$$

8.1.4 Haromic mean (H.M):

Harmonic mean is the reciprocal of the arithmetic mean of the given observations :

(a) In Case of Ungrouped Data :If $X_1, X_2, X_3, \ldots\ldots X_n$ is a given set of *n* observations then their harmonic mean is :

$$H = \frac{n}{\Sigma \frac{1}{x}}$$

(b) In Case of Grouped Data : In case of frequency distribution the harmonic mean is given by the formula:

$$H = \frac{n}{\Sigma \frac{f}{x}}$$

Where, f= total frequency

x= value of variable

8.1.4.1 Merits of Harmonic Mean:

1. It is based on all observations.
2. It is not affected much by fluctuations of sampling.
3. As reciprocal values are involved, it gives greater weightage to smaller observations.

8.1.4.2 Demerits of Harmonic Mean :

1. It is difficult to understand and calculate for biologists.
2. Its value cannot be obtained if any one of the observations is zero.

Example 1**:** The following table gives the weight of 31 persons in a sample enquiry. Calculate the mean weight using Geometric and Harmonic Means.

Weight(X) :	130	135	140	145	146	148	149	150	157
No. of persons (f):	3	4	6	6	3	5	2	1	1

$$G.M. = \text{Antilog } \frac{1}{n} f \log X = \text{Antilog } \frac{1}{31} \times 66.7710$$

$$= \text{Antilog } 2.15390$$

Therefore, $G.M = \text{Antilog } 2.15390 \quad = 142.5$ **Ans.**

$$H.M. = \frac{N}{\Sigma f/x} = \frac{31}{0.2177} = 142.39$$

Here $N = 31$ and $\Sigma f/x = 0.2177$

Ans: (1) Mean weight according to G.M. method = 142.5

(2) Mean weight according to H.M. method = 142.39

Relation between arithmetic mean (AM), geometric mean(GM) and harmonic mean(HM).

1. When Observations are Equal

If in a set all the observations are equal, then

A.M. = G.M. = H.M.

Example 2 The two positive items are equal and their measures are 6 and then their

1. Arithmetic Mean (A. M.) = $\frac{6.+6}{2} = 6$

2. Geometric Mean (G.M.) = $\sqrt{6x6 = 6}$

3. Harmonic Mean (H.M.) = $\frac{2}{\frac{1}{6}+\frac{1}{6}} = \frac{2}{\frac{1}{3}} = \frac{2x3}{1} = 6$

2. When Observations are Unequal

When in a set of observations, the size of observations varies, the arithmetic mean (A.M.) is greater than GM and GM is greater than HM.

A.M. > G.M. > H.M.

or X > G.M. > H.M.

Example 3**:** The two positive items are 4 and 9. In this case

1. Arithmetic Mean (A.M.) = $\frac{4+9}{2} = 6.5$

2. Geometric Mean (G.M.) = $\sqrt{4x9 = 6}$

3. Harmonic Mean (H.M.) = $\frac{2}{\frac{1}{4}+\frac{1}{9}} = \frac{2}{\frac{13}{36}} = \frac{2x36}{13} = 5.5$

A.M. > G.M. > H.M. = 6.5 > 6 > 5.5

8.2 Median (md or m_d):

If the values of a variable are arranged in ascending or descending order of magnitude, the median is that value which divides the whole data into two equal parts, one part having all values smaller than the median value and other part having all the values greater than the median value. In other words, 50% of the observations will be smaller than the median while the other 50% will be larger than the median. It means the value of the middle observation or the mean value of two middle observations is called median. As a matter of fact median is called the positional average. Median is calculated differently for ungrouped and grouped data.

8.2.1 Median of Ungrouped Data:

To calculate median of ungrouped data, the values of a variable (i.e. all the observations about the variable) are arranged in the order of magnitude either in ascending or descending order. The middle-most value in this arrangement represents the median.

Example 1**:** To calculate the median of following seven observations:

100	97	110	200	75	120	150

Solution:

Step 1. The observations of the raw data are arranged in ascending order of magnitude in the sequence:

75 97 100 110 120 150 200

Step 2. Total number of observations is found out. It is 7 in this case.

Step 3. The middle most value in the above series is 110. This is the median.

Therefore, the median of above series is 110.

Calculation of Median:

Median can be calculated in the following ways:

1. Calculation of Median when Number of Observations is Odd: When number of observations

(n) is uneven the median is calculated by using the following formula:

$$\text{Median} = \frac{\text{No of observations+1}}{2} = \left[\frac{n+1}{2}\right]_{th} \text{observations}$$

In the above example:
(a) the number of observations is 7

(b) Their median will be $\frac{7+1}{2} = 4 = 4$ th position

(c) In the series fourth position is occupied by 110.
(d) Therefore, median of the above series is 110.

2. Calculation of Median when Number of Observation (n) is Even: When number of observations is even, there is no unique median. The median in such cases is located half way between the two middle items. Therefore, the median is taken as the arithmetic mean of nth and $(n + 1)^{th}$ observations.

Example 2**:** To calculate median of following 6 observations:

75	**97**	**100**	**120**	**150**	**175**

Step 1. Arithmetic mean of 6 observations is $\frac{n}{2} = \frac{6}{2} = 3$
= third observation

Step 2. Arithmetic mean of 6 observations $\frac{n}{2} = \frac{6}{2} = 3$
= third observation

Step 3. Arithmetic mean of 3rd + 4th observations = Median of 6 observations

(a) 3rd observation = 100

(b) 4th observation = 120

Arithmetic mean = $\frac{100+120}{2} = \frac{220}{2} = 110 = \text{Median}$

Ans. Median = 110

Example 3**:** The number of patients that visited a doctor for consultation for 10 consecutive days is arranged in an increasing order in the following table. Find out the median number of the patients that visited the doctor per day.

8	**10**	**12**	**14**	**16**	**18**	**19**	**20**	**22**	**25**

Solution:

Step 1. The number of observations {i.e. the number of days) n = 10

Step 2. Arithmetic mean of 10 observations $\frac{10}{2} = 5$

= 5th observation

Step 3. Arithmetic mean of 10 observations $+1 = \frac{10}{2} + 1 = 6$

= 6th observation

Step 4. Median for above 10 observations = Arithmetic mean of 5th value and of 6th value.

(a) 5^{th} observations =16

(b) 6^{th} observations =18

Median = 17 Ans.

3. Calculation of Median for Simple Frequency Distribution: For calculation of median for frequency distribution of ungrouped data, cumulative frequency corresponding each value of variable is calculated.

The value of the variable corresponding to the cumulative frequency of $\frac{N+1}{2}$ is called medial where N is the total frequency.

Example 4**:** Calculate the median value of the following variables on the basis of following simple frequency distribution.

Variable (X_1)	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Frequency (f_1)	**1**	**4**	**12**	**9**	**2**	**1**	**1**

Solution:

Variable (X_1)	**Frequency (f_1)**	**Cumulative Frequency (F)**
1	1	1 = (0+1)
2	4	5 = (4+1)
3	12	17 = (12+5)

4	9	26 = (9+17)
5	2	28 = (2+26)
6	1	29 = (1+28)
7	1	30 = (1+29)

Where total number of frequencies (N) = 30

Median value of variable = $\frac{N+1}{2} = \frac{30+1}{2} = 15.5$

8.2.2 Median of Grouped Data:

The median of grouped data is calculated by the following formula
L_1 = The lower limit of that class interval where median falls.
N or Σf= Total number of frequencies

f= Frequency

F = The cumulative frequency just above that class interval where median falls.

fm = The frequency of that class interval where median falls.

i = The width of the class interval.

Median has wirier application in behavioral biology and environmental pollution. Etc

Solution:
Here Median $\frac{\Sigma f}{2} = \frac{76}{2} = 38$

The class interval in which 33'" item (in cumulative frequency) falls is 15-19. See table given above The exact limits or class boundary of this interval are

14.5-19.5. The lower limit of this class interval (L₁): L_1 = 14.5.

F=13, fm = 13x2 = 26, and i =5

$$\text{Median} = L_1 + \frac{\left[\frac{\Sigma f}{2} - F\right]}{\text{fm}} X_1$$

$$= 14.5 + \frac{38-13}{26} \times 5 = 14.5 + \frac{25}{26} \times 5 = 19.31$$

The median of the above data is 19.31 Ans.

8.2.3 Merits of Median:

1. It is rigidly defined.

2. Median is easy to understand and easy to calculate.

3. Median is not affected by extreme observations and is very useful in the case of skewed distribution. .

4. Median can be computed while dealing with a distribution with open end class.

5. Median S best for qualitative data.

8.2.4 Demerits of Median:

1. Median cannot be determined in the case of even number of observations. We merely estimate it as the arithmetic mean of the two middle terms.

2. Median is relatively less stable than means, particularly for small samples since it is affectec more by fluctuations of sampling as compared with arithmetic mean.

3. Median is a positional average. It cannot be accepted for each and every observation.

4. It cannot be subjected to algebraic treatment.

Table Formulae of calculation of median at a glance

8.3 MODE (MO or M_0):

Mode is the most frequently occurring value in a data. It means for a given data, mode may or may not exist. For example, let's observe mode for the following 3 sets of data :

(a) 10, 10, 9, 8, 5, 4, 12, 10 : One mode. i.e., 10

(b) 10, 10, 9, 9, 12, 15, 5 : Two modes i.e., 10 and 9

(c) 4,6,7, 15, 12,13,10 : No mode.

We note that first set of data (a) has single mode 10, the second set of ata (b) has two modes 10 and 9 and the third set of data (c) has no mode. Set (b) has 10 and 9 as modes because they both occur 2 times and they occur more often than other values.

In terms of frequency distribution, mode is the variable at which the curve peaks. Thus, based on the number of peaks in the curve the frequency distribution may be of following three types:

1. Unimodal Frequency Distribution: A distribution data having one mode is called unimodal frequency distribution.

2. Bimodal Frequency Distribution : The frequency distribution having two peaks is called bimodal frequency distribution.

3. Multimodal Frequency Distribution: The frequency distribution with more than two peaks is called the multimodal frequency distribution.

4. Antimode: In U-shaped distribution the low point at the middle of the distribution is known as an antimode.

Definition of Mode and modal class :

According to A.U. Tuttle, 'Mode is that value in a sample or data which has the greatest or largest frequency density in the frequency table'. The class having greatest frequency is called modal class. The modal class can be determined by inspection but the actual value of mode will lie in the class interval and may not be at the midpoint of the class.

8.3.1 Calculation of Mode:

8.3.1.1. Calculation of Mode of individual Series or Ungrouped Data:

It can be computed either by inspection or by frequency distribution.

1. Calculation of Mode by Inspection;
2. Calculation of Mode by Frequency Distribution

In this method the data is arranged in increasing order. It is then observed that how many times each value in the data is repeated. The item or value which occurs most frequently represents the mode of that data.

Example 1 Calculation of Mode by Inspection

Variable x	**32**	**22**	**29**	**25**	**17**	**25**	**40**

Solution:
Step1. Arrange the data in increasing order, i.e.

Variable X: 17 22 25 25 29 32 40

Step2. Value 25 of X in this series has occurred twice while all others are represented just once. Therefore, mode of this data is 25.

8.3.1.2. Calculation of Mode by Frequency Distribution:

When the number of items in a series is very large, individual items are converted into frequeni distribution. The mode is then calculated as the value corresponding to the highest frequency.

Example 1**:** Calculation Mode on the basis of simple frequency distribution of a variable

variable **(x) Number of flowers**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Frequency (f) (Plants	**1**	**4**	**12**	**9**	**2**	**1**	**1**

Solution:

1. In this example, the number of flowers on 30 plants varies from 1 to 7. This is called variable

2. Plants bearing 3 flowers are maximum, i.e. 12 in number. This represents their frequency.

3. 3 is the value with highest frequency of 12.

Therefore, Mode = 3 flowers per plant.

3. Calculation of Mode of Discrete Series**:**

In discrete series, mode can be determined by inspection method or by grouping method The inspection method is similar to the one discussed in ungrouped data.

4. Calculation of Mode of Discrete Series by Grouping Method

This method is used in cases where there is regularity and homogeneity in the series. Inspectiol method is usually not reliable and grouping method is applied. It involves preparation of grouping table as follows:

Example 2**:** from the given data calculate the value of mode by grouping method

Marks(X) :	1	2	3	4	5	6	7	8	9	10	11	12
No. of Students (f):	4	7	8	12	16	14	9	7	17	5	3	2

1. In the column I of the grouping table, the values of variable are

   arranged in the ascending order.
2. In column 2, corresponding frequencies are written.

3. The frequencies are grouped in two's beginning with the first value.

4. In column 3, the frequencies are grouped in two's beginning with

   second value of the series.

5. In column 4, the frequencies are grouped by three values (i.e. 1,2 and

   3) starting with the first value. Write them as shown in the table.

6. In column 5, the frequencies are grouped into three values beginning with the third value.

7. Underline the maximum grouped frequency in each column.

8. Prepare the analysis table showing the items corresponding to the maximum frequencies I different columns of the grouping table. In analysis table column numbers 1 to 6 showin frequencies are placed on the left hand side and sizes of the items (1 to 12) on the too. The

9. From the analysis table we find out the variable which has maximum frequency of distribution.

Result: From the analysis table it is evident that value 5 or item 5 has maximum frequency.

8.3.1.3. Calculation of Mode in a Continuous Series

Determination of mode in case of continuous frequency distribution is complicated and involves

1. First of all, the modal class is ascertained either by inspection method or by grouping method.

2. After determining the modal class, the exact value of mode is calculated by using the following formula

where Li = Lower limit (boundary) of modal class
frn = Frequency of modal class or the maximum frequency

$$\text{Mode (Z)} = L_1 + \left(\frac{fm - f_1}{2fm - (f_1 - f_2)}\right) \times C$$

f1 = frequency of class just preceding the modal class

f2 = frequency of class just succeeding the modal class

C = Class interval or class width of the classes.

Example 1: In a class following is the distribution of marks of 85 students. Calculate the modal class and mode of the followina data:

Solution:

Marks(Group data)	20-25	20-30	30-35	35-40	40-45	45-50	50-55	55-60
No. of Students (f):	5	7	8	18	25	12	7	5

Modal class by inspection is 40-45

Fm = 25 fo = 12 Lower limit of modal class L = 40

f_1 = 18

Class interval C = 5

Putting these values in the formula

$$\text{Mode (Z)} = L_1 + \left[\frac{fm - f_1}{2fm - (f_1 - f_2)} \times C\right]$$

$$= 40 + \frac{25 - 18}{(2 \times 25) - (18 - 12)} \times 5$$

$$= 40 + \frac{7}{50 - 30} \times 5$$

$$= 40 + \frac{7}{20} \times 5 = 40 + \frac{7}{4} = \frac{160 + 7}{4}$$

$$= \frac{167}{4} = 41.75$$

Modal Class has value = 41.75 Ans.

8.3.2 Merits of Mode:

1. Mode is easy to calculate and understand.

2. It is not affected .by extreme observations and as such is preferred to arithmetic mean while dealing with extreme observations.

3. Mode can be calculated from a grouped frequency distribution with open-end classes.

8.3.3 Demerits of Mode:

1. Mode is not rigidly defined. It is ill defined if the maximum frequency is repeated or occurs in the very beginning or at the end of the distribution.

2. As compared to mean, mode is affected to a great extent by the fluctuations of sampling.

3. It is not suitable for algebraic treatment.

Relationship between Mean (m), Median (md) and Mode (mo):

In case of a symmetrical (normal) distribution mean, median and mode have the same value. Therefore, in a graph all three of them fall at the same position. It means the three values coincide (see Fig. 30) i.e.. Mean = Median = Mode.

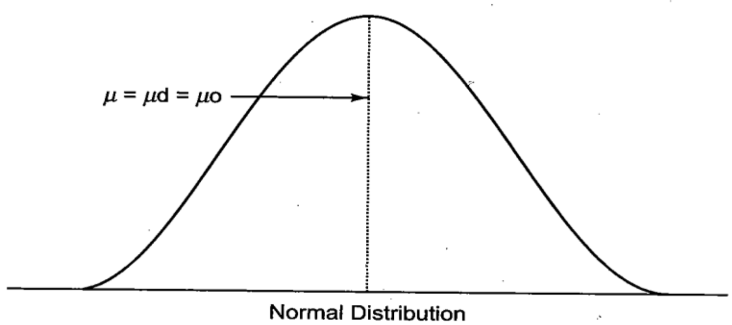


Fig. 3 Normal distribution.

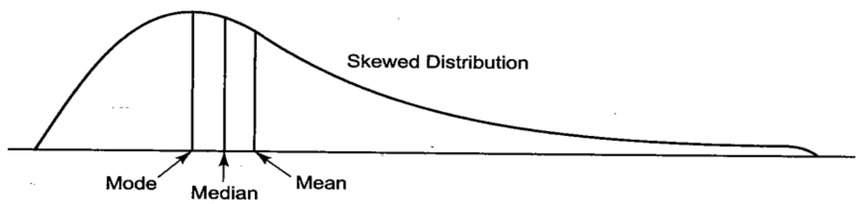


Fig. 4. Diagram showing relation between mean, median and mode in case of skewed distribution.

Fig.30 : Relationship between Mean, Median and Mode.

But in skewed distribution, mean and mode usually lie on the two ends and median lies between them and in that case (Fig.30).

1. For a positively skewed distribution, mean will be greater than median and median is greater than mode, i.e.,

$$\mu > Md > Mo \quad \text{or} \quad Mo < Md < \mu$$

2. For a negatively skewed distribution the order of the magnitude of the three averages will be reversed, i.e.,

$$Mo > Md > \mu \text{ or } \mu < Md < Mo$$

8.4 Dispersion :

Introduction to range and variance:

- In general, Range means the value we obtained as a result of difference between a greater value and a smaller value.
- The variance of a random variable or distribution is the expectation, or mean, of the deviation squared of that variable from its expected value or mean. (Source - Wikipedia)
- In this article of *range variance*, we are going to discuss about range and variance of the data set.

8.4.1 Range:

Steps involved to find range:

Step 1: First arrange the numbers from ascending order to descending order.

Step 2: Identify the larger value in the given data

Step 3: Identify the smaller value in the given data

Step 4: Find the difference between larger and smaller value.

Example problems:

Example 1:

Find the range for the following data set:

{ 45 , 15 , 16 , 25 , 10 }

Solution:

Range = larger value - smaller value

= 45 - 10

= 35

Example 2:

Find the range for the following data set:

{ 21 , 30 , 19 , 32 , 36 }

Solution:

Range = larger value - smaller value

= 36 - 19

= 17

Practice problems:

1) Find the range for the following data set:

{ 23 , 15 , 20 , 31 , 44 }

2) Find the range for the following data set:

{ 10 , 56, 29, 56 , 34}

Answer key:

1) 29 2) 46

8.4.2 Variance:

Steps involved to find Variance:

Step 1: Mean:Find the average or mean value of the given set of data.

Step 2: Variance:To find the variance, get each difference from the mean value and square the each value and finally average the result.

Example 1: Calculate the variance of the numbers 2, 7 and 9.

$$\bar{X} = \text{Mean} = \frac{\Sigma X}{N} = \frac{18}{3} = 6 \quad \text{Therefore,} \quad \bar{X} = 6,$$

$$V = \text{Variance} = \frac{\Sigma(X - \bar{X})^2}{N} = \frac{26}{3} = 8.76. \quad \textbf{Ans.}$$

Example 2: Find the variance of the following data set:

{ 40, 20, 30, 50 }

Solution:

Step 1: Mean

$$\text{Mean} = \frac{40+20+30+50}{4}$$

$$= \frac{140}{4}$$

$$= 35$$

Step 2: Variance

Variance =

=

$$= \frac{25+225+25+225}{4}$$

$$= \frac{500}{4}$$

= **125**

Practice problems:

1) Find the variance of the data set: { 40, 25, 20, 47, 63 }

Answer: Variance = 239.6

2) Find the variance of the data set: { 33, 46, 57, 25 }

Answer: Variance = 149.69

3) Find the variance of the data set: { 66, 26, 96, 79 }

Answer: Variance = 666.69

8.5 Standard Deviation (SD):

Standard deviation of a series is the positive square root of the arithmetic mean of the squares of deviations of the various items from the arithmetic mean of the series. It is also called root mean square deviation. It is represented by Greek symbol (σ) and in short form by S.D.

8.5.1 Computation of Standard Deviation from Ungrouped Data:

Standard deviation is calculated by two methods in ungrouped data:

1. Indirect method

2. Direct method.

1. Indirect Method**:** For indirect method deviation is obtained from mean. The formula used is:

$$\sigma = \sqrt{\frac{\Sigma X^2}{N}} \quad \text{or} \quad \sqrt{\frac{\Sigma dx^2}{N}} \quad \text{and} \quad \sqrt{\frac{\Sigma X^2}{N-1}} \quad \text{or} \quad \sqrt{\frac{\Sigma dx^2}{N-1}}$$

where, X or dX = deviation obtained from actual mean.

N = Total number of observations.

If size of sample is small, i.e., the number of observations is less than 30, standard deviation is computed by using N-I in the denominator of above formula in place of N. If the size of sample is more than 30 then formula used will be :

$$\sigma = \sqrt{\frac{\Sigma X^2}{N}} \quad \text{or} \quad \sqrt{\frac{\Sigma dx^2}{N}}$$

Calculate standard deviation from the given data by indirect method:

From the above data we conclude that:

1. Mean a= 32

2. Deviation Zrfx= 10

3. Number of observations in above data N = 10

4. Deviation square £0~ = 771

$$\bar{X} = a + \frac{\Sigma dx}{N}$$

$$= 32 + \frac{10}{10} = 32 + 1 = 33$$

$$\sigma = \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2} = \sqrt{\frac{771}{10} - \left(\frac{10}{10}\right)^2}$$

$$= \sqrt{\frac{771}{10} - 1} = \sqrt{\frac{761}{10}} = \sqrt{(77.6)} = 8.7 \quad \textbf{Ans.}$$

2. Direct Method: In direct method there is no need to obtain actual mean. Deviate obtained from assumed mean. Small x prime (xQ is used for deviation obtained from assi mean. Following formula is used to obtain standard deviation by direct method in ungrouped data:

$$\sigma = \sqrt{\frac{\Sigma x'^2 - \frac{(\Sigma x')^2}{N}}{N-1}}$$

Computation of standard deviation by direct method is carried out by using assumed mea involves following six steps in a fixed order:

Step 1. Find mean of the series.

Step 2. Find each deviation from the mean (X - X).

Step 3. Square each deviation, finding X^2.

Step 4. Sum the squared deviations, finding Σx^2

Step 5. Divide this sum by N or N-1, finding

$$\frac{\Sigma x^2}{N} \quad \text{or} \quad \frac{\Sigma x^2}{N-1}$$

Step 6. Extract the square root of the result of step 5. This is standard deviation.

Calculate standard deviation from the data given in above table by direct

method.

Solution:

$$\sigma = \sqrt{\frac{\Sigma dx^2}{N}} = \sqrt{\frac{761}{10}}$$

$$= \sqrt{(76.1)} = 8.729$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{330}{10} = 33 \quad \textbf{Ans.}$$

8.5.2 Computation of Standard Deviation from Grouped Data:

There are two methods to compute standard deviation from available data: Long method and short method.

1. Long Method : Following formula is used to obtain standard deviation by long method :

$$\text{S.D.} \quad \sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f}} \quad \text{(S.D: standard deviation)}$$

Steps in calculating standard deviation from grouped data by long method involves following steps:

Step 1. Find midpoint of each class interval.

Step 2. Find mean value of the series using formula: $\frac{\Sigma f.x}{f}$

Step 3. Find each deviation from the mean.

Step 4. Square each deviation, finding X^2

Step 5. Multiply each squared deviation with corresponding frequency, finding fxx^2

Step 6. Sum the squared deviation multiplied by frequency, finding Σfxx^2

2. Short Method: Following formula is used to calculate standard deviation by short method :

$$\text{S.D. or} \quad \sigma = i \cdot \sqrt{\frac{\Sigma f \cdot x'^2 - c^2}{\Sigma f}}$$

where, x! = Deviation calculated from assumed mean
c = Correction i = Class interval

Methods of Calculating Standard Deviation at a Glance

Type of Series	Method of Calculation
1. Ungrouped Data (Individual series)	S.D $(\sigma) = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{\Sigma dx^2}{N}}$ (Standard deviation)
2. Grouped Data (Discrete Series)	S.D $(\sigma) = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}}$
3. Grouped Data (Continuous series)	Long method 1. $\sigma = \sqrt{\frac{\Sigma f \cdot X^2}{\Sigma f}}$ 2. Short method $\sigma = i\sqrt{\frac{\Sigma f \cdot X^2 - c^2}{\Sigma f}}$

Coefficient of Standard Deviation

For comparative study, coefficient of standard deviation is obtained by the following formula:

$$\text{Coefficient of standard deviation} = \frac{\text{Standard deviation}}{\text{Arithmetic mean}} = \frac{\text{S.D.}}{\bar{X}}$$

Merits of Standard Deviation

1. It summarizes the deviation of a large distribution from mean in one figure used as a unit of variation.

2. It indicates whether the variation of difference of an individual from the mean is real or by chance.

3. It helps in calculating the standard error.

4. It helps in finding the suitable size of sample for valid conclusions.

Demerits of Standard Deviation

1. It gives weight age to only extreme values.

2. The process of squaring deviations and then taking square root involves lengthy calculation.

Significance of Standard Deviation

1. Standard deviation is based on all the observations.

2. The squaring of the deviations $(X - X)^2$ removes the drawbacks of ignoring the signs of deviations in computing the mean deviation.

3. Of all the measures of dispersion, standard deviation is affected least by fluctuations of sampling.

.

8.6 Standard Error (SE):

When more than one samples are drawn from a population, the standard deviation for each mear will be found to differ slightly. In theory, the mean of different samples drawn from the same population must be the same or very close to each other so that the single sample is a reliable true measure of the population. But in practice this is not' so and means of samples show variations.

The square root of the variance of the sampling distribution $\sqrt{\sigma \frac{-2}{x} or \frac{\sigma}{\sqrt{n}}}$ is called the standard error of the mean or simply standard error.

$$SE = \frac{S \text{ or } SD}{\sqrt{(n)}} \quad \text{or} \quad \frac{\sqrt{S^2 \text{ or } SD}}{(n)}$$

Significance of Standard Error
Standard error indicates that

1. a population with a large variance (02) will provide a large variance in all its sample means.
2. by increasing sample size (n), the variance in sample means (X) can be deduced.
3. the variance of original population is usually more than the variance of sampling distribution.

8.7 Coefficient of Variation or Coefficient of Variance :

It is measurement of relative dispersion. It gives an idea about the extent to which varieties are scattered around their central value. Therefore, two distributions having same central values can be compared directly with the help of various measures of dispersion. When we calculate standard deviation of two populations having different mean, the standard deviation alone does not give us correct comparative idea about the extent of variability in two populations. If this standard deviation is expressed as percentage of mean, such a parameter/statistics provides a comparative idea of the extent of variability. This parameter is called co-efficient of variability and can be calculated as follows :

$$\text{Coefficient of variability (C.V)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma \times 100}{\bar{X}}$$

The coefficient of variation remains unaltered by a change in scale, for example, from feet centimeter, but it is altered by a change of origin which affects the mean but does not alter the standard deviation (S.D).

Example 1:Given mean 20 and 16, find the coefficient of variation.

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100$$

$$\text{or} \quad \text{C.V.} = \frac{16}{20} \times 100 = 80\% \textbf{ Ans.}$$

STUDY QUESTIONS:

Short Questions

1. **Define Mode**
2. **What is a Range?**
3. **Define Standard Deviation(SD)**
4. **What is Coefficient of Variation?**

Long Questions

1. **Write about the various types of Mean.**
2. **Describe the computation of Standard Deviation for Grouped and Ungrouped data.**
3. **Explain the concept of standard error.**

UNIT IX

Structure

9.1 Probability

9.2 Hypothesis Testing

 9.2.1 Null Hypothesis

9.3 Normal Distribution

9.4 Confidence Interval

9.5 P Value

9.1 Probability:

Probability Values A probability must fall in the range 0.00 - 1.00 . If the probability of event A = 0.00, then A is certain not to occur. If the probability of event A = 1.00, then A is certain to occur. Every event has a complement: For example, the complement of event A is the event not A, which is usually symbolized as A . The probability of an event plus the probability of its complement must be equal to 1. This is just the same thing as saying that the event must either happen or not happen. A priori probability is a priori method of computing probability is also known as the classical method. It might help to think of it as the expected probability value (e.g., like expected frequencies used in calculating the chi-squared statistic). number of events classifiable as total number of possible events.

The a posteriori method is sometimes called the empirical method. Whereas the a priori method corresponds to expected frequencies, the empirical method corresponds to observed frequencies. number of times has occurred total number of events

Example 1: Consider a fair six-sided die (die = singular of dice). What is the probability of rolling a 6? According to the a priori method, p(6) = 1/6. But to compute p(6) according to the empirical method, we would have to roll the die some number of times (preferably a large number), count the number of sixes, and divide by the number of rolls. As alluded to earlier, statistics like chi-squared involve comparison of a priori and empirical probabilities.

Two reasons why probability is important for the analysis of linguistic data: Joint and conditional probabilities are used to analyze corpus data Probability plays an important role in statistical hypothesis testing

Simple probability:

If you toss a dice with six number (i.e. 1,2,3,4,5,6) what is the probability that you will toss a 6?

P(6) = 1/6 = 0.1666

• Probability values range from 0 to 1.

• Adding all probabilities of the sample yields 1.

• If two events are independent, the probability is the sum of their individual probabilities. Probability

• Two events A and B are independent if knowing that the occurrence of A does not change the probability of the occurrence of B.

9.2 Hypothesis Testing:

The main purpose of statistics is to test a hypothesis. For example, you might run an experiment and find that a certain drug is effective at treating headaches. But if you can't repeat that experiment, no one will take your results seriously. A good example of this was the cold fusion discovery, which petered into obscurity because no one was able to duplicate the results.

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

Hypothesis testing can be one of the most confusing aspects for students, mostly because before you can even perform a test, you have to know what your null hypothesis is.
Often, those tricky word problems that you are faced with can be difficult to decipher. But it's easier than you think; all you need to do is

1. Figure out your null hypothesis,
2. State your null hypothesis,
3. Choose what kind of test you need to perform,
4. Either support or reject the null hypothesis.

9.2.1 Null Hypothesis:

If you trace back the history of science, the null hypothesis is always the accepted fact. Simple examples of null hypotheses that are generally accepted as being true are: DNA is shaped like a double helix. There are 8 planets in the solar system (excluding Pluto). Taking Vioxx can increase your risk of heart problems (a drug now taken off the market)

Rejecting the null hypothesis

Ten years ago, we believed that there were 9 planets in the solar system. Pluto was demoted as a planet in 2006. The null hypothesis of "Pluto is a planet" was replaced by "Pluto is not a planet." Of course, rejecting the null hypothesis isn't always that easy — the hard part is usually figuring out what your null hypothesis is in the first place.

Example 1: A sample of 200 people has a mean age of 21 with a population standard deviation (σ) of 5. Test the hypothesis that the population mean is 18.9 at α = 0.05.

Step 1: State the null hypothesis. In this case, the null hypothesis is that the population mean is 18.9, so we write:
H_0: μ = 18.9

Step 2: State the alternative hypothesis. We want to know if our sample, which has a mean of 21 instead of 18.9, really is different from the population, therefore our alternate hypothesis:
H_1: μ ≠ 18.9

Step 3: Press Stat then press the **right arrow** twice to select TESTS.

Step 4: Press 1 to select **1:Z-Test…**. Press ENTER.

Step 5: Use the **right arrow** to select **Stats**.

Step 6: Enter the data from the problem:
μ_0: 18.9
σ: 5
x: 21
n: 200
μ: $\neq\mu_0$

Step 7: Arrow down to **Calculate** and press ENTER. The calculator shows the p-value:
$p = 2.87 \times 10^{-9}$

This is smaller than our alpha value of .05. That means we should **reject the null hypothesis**.

9.3 Normal distribution :

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. The normal distribution is the most common type of distribution assumed in statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations. The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution (Fig.31).

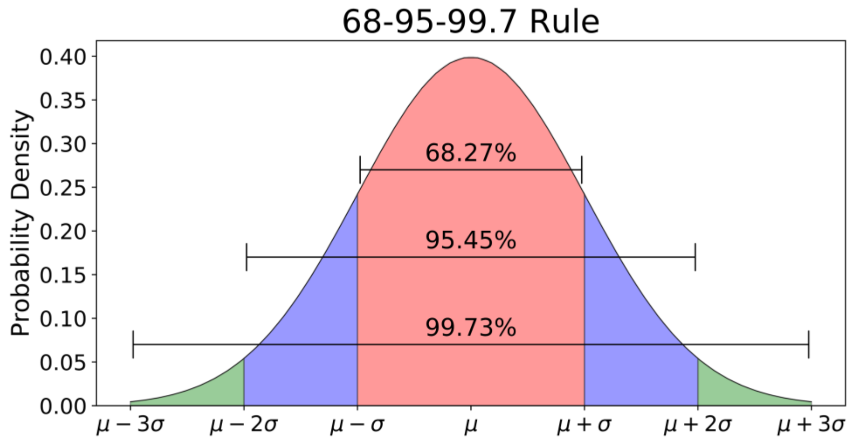


Fig.31 : Normal Distribution

The center of the curve represents the mean, median, and mode.

• The curve is symmetrical around the mean.

• The tails meet the x-axis in infinity.

• The curve is bell-shaped.

• The total under the curve is equal to 1 (by definition).

9.4 Confidence Interval:

The confidence intervals account for that margin of error. To do this, we'll use the same tools that we've been using to understand hypothesis tests. I'll create a sampling distribution using probability distribution plots, the t-distribution, and the variability in our data. We'll base our confidence interval on the energy cost data set that we've been using. When we looked at significance levels, the graphs displayed a sampling distribution centered on the null hypothesis value, and the outer 5% of the distribution was shaded. For confidence intervals, we need to shift the sampling distribution so that it is centered on the sample mean and shade the middle 95%.

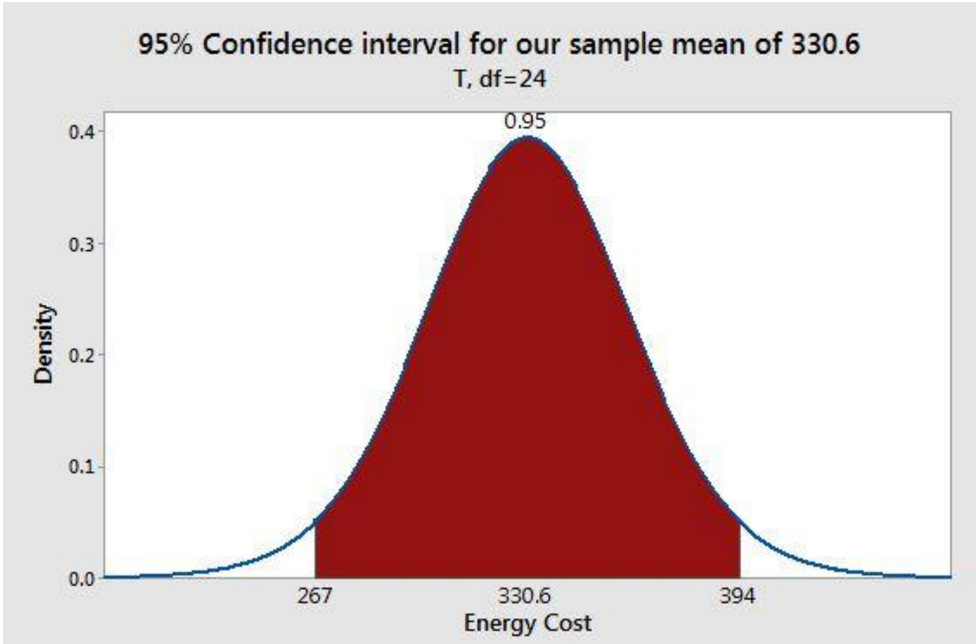


Fig.32 :Confidence Interval

The shaded area shows the range of sample means that you'd obtain 95% of the time using our sample mean as the point estimate of the population mean. This range [267 394] is our 95% confidence interval. Using the graph, it's easier to understand how a specific confidence interval represents the margin of error, or the amount of uncertainty, around the point estimate. The sample mean is the most likely value for the population mean given the information that we have. However, the graph shows it would not be unusual at all for other random samples drawn from the same population to obtain different sample means within the shaded area. These other likely sample means all suggest different values for the population mean. Hence, the interval represents the inherent uncertainty that comes with using sample data. You can use these graphs to calculate probabilities for specific values. However, notice that you can't place the population mean on the graph because that value is unknown.

9.5. P Value:

It's good science to let people know if your study results are solid, or if they could have happened by chance. The usual way of doing this is to test your results with a p-value. A p value is a number that you get by running a hypothesis test on your data. A P value of 0.05 (5%) or less is usually enough to claim that your results are repeatable. However, there's another way to test the validity of your results: Bayesian Hypothesis testing. This type of testing gives you another way to test the strength of your results. The p-value indicates the probability that we will obtain the distribution in our sample given that there is no relationship between x and y in the true population. The p-value is conditional on the hypothesis that the null hypothesis is true.

If there is no relationship (difference) between X and Y in the true population, then there is a less than a 5% chance (i.e. 1 out of 20 chance) that we obtain the distribution in a given sample.

STUDY QUESTIONS:

Short Questions

1. **Define Probability**
2. **Define Null Hypothesis**
3. **What is Normal Distribution?**
4. **What is a P Value?**

Long Questions

1. **Give an Account on Hypothesis testing**
2. **Enumerate the characteristics features of a normal distribution?**
3. **Explain the role of confidence interval.**

UNIT X

Structure

10.1 Common Statistical Tools:

10.1.1 Chi-square test (χ2) :

Introduction:

Standard error, student's r-test, Z-test, F-test, etc., discussed in the earlier chapters are parametric tests. They are applied to quantitative data and are based upon an assumption about the distribution of certain parameters. The nonparametric tests, on the other hand, are not concerned with any population distribution. Chi-square test is the most commonly used nonparametric test in biological experiments. It is computed on the basis of frequencies in a sample and is applied only for qualitative data such as for intelligence, colour, immunity, health, response to drug, etc. Chi-square test is used as a test of significance when data is expressed in frequencies or in terms of percentages. Chi- square test enables us to determine the degree of deviation between observed frequencies and the theoretical frequencies and to conclude whether the deviation between observed frequencies and expected (=theoretical) frequencies is due to

error of sampling or due to chance. Chi-square is a mathematical expression, representing the ratio between experimentally obtained or observed results (0) and the theoretically expected results based on certain hypothesis. Chi-square test is a test of significance. It uses data in the form of frequencies (i.e. the number of occurrence of an event). Chi-square is calculated by dividing the overall deviance square in the observed and expected frequencies by the expected frequency.

Formula for determining ($\chi 2$)

The chi-square **(χ2)** is calculated by using following formula:

$$x2 = \Sigma \frac{(O - E2)}{E}$$

The alternative formula is :

$$x2 = \Sigma \frac{(O - E2)}{E} - (N),$$

where,

0 = observed frequency in a class
E = expected frequency in a class

2 = summation over all classes

N= total number of observations
From these equations, the value of **(χ2)** will be zero, if O = E in each class. But due to chance error this never happens. However, the observed results are based on the same principal from which expected frequencies are calculated.

The values of **(χ2)** depend on the number of classes, i.e., on the number of degrees of freedom (df) and the critical level of probability (5% or 1%). The expected value of 2 is obtained from the chi-square table. This expected value can be compared with the value calculated from the data. If the expected value is lower than the observed or computed value, the hypothesis is rejected. Suppose out of 100 participants participating in an awareness program camp, there were 60 girls and 40 boys. These results refer to observed frequencies and are denoted by O. According to Null hypothesis, the number of boys and girls participating in the programme should be equal, i.e,, the expected number of boys and girls participating in the awareness programme should be 50 each. These frequencies are theoretical or expected frequencies. They are denoted by E. The observed frequencies are 40 boys and 60 girls. These are represented by 0. All these frequencies can be represented in the following contingency table:

The above results are used to test the hypothesis using chi-square formula as follows:

$$= \frac{(40-50)^2}{50} + \frac{(60-50)^2}{50} = \frac{(10)^2}{50} + \frac{(10)^2}{50}$$

$$= \frac{100}{50} + \frac{100}{50} = 2 + 2$$

$$\chi^2 = 4$$

(χ2) value for girls and boys will be:

$$\chi^2 = \left[\frac{(40)^2}{50} + \frac{(60)^2}{50}\right] - 100 = \left[\frac{1600}{50} + \frac{3600}{50}\right] - 100$$

$$= \left[\frac{1600 + 3600}{50}\right] - 100 = \frac{5200}{50} - \frac{100}{1} = \frac{5200 - 5000}{50} = \frac{200}{50}$$

$$\chi^2 = 4$$

In the above example, computed XZ value is 4. From the chi-square table the expected value of **(χ2)** for I degree of freedom is 3.841. It means observed or computed value af 2 is greater than expected value of **(χ2)** (4 > 3.841). Therefore the hypothesis that there is no difference between expected and observed values, is rejected.

Alternate formula :

Chi-square can be determined using an alternative formula given below :

Chi- square or **(χ2)** $\Sigma \frac{O2}{E} - N$

By using the above data in this equation:

$$X2\left[\frac{(40)2}{50} + \frac{(60)2}{50}\right] - 100 = \left[\frac{1600}{50} + \frac{3600}{50}\right] - 100$$

$$= \left[\frac{1600 + 3600}{50}\right] - 100 = \frac{5200}{50} - \frac{100}{1} = \frac{5200 - 5000}{50} = \frac{200}{50}$$

Chi-square distribution:

Chi-square distribution is a continuous distribution. Its probability density function is given

$$P(x)^2 = y^0\ (x^2)^{1/2\ (v-2)}\ e\frac{-1}{2}X2$$

X^2 Ci-suquare
f = Constant depending on degree of freedom

v=Degreeoffreedom=(n-1).

Characteristics of chi-square of chi-square distribution :

1. Chi-square curve is always positively skewed, i.e. **χ2** value is always positive.

2. Chi-square values increase with the increase in degree of freedom.

3. The standard deviation of $\chi 2$ distribution is equal to √2v where v is the degree of freedom.

4. The mean of distribution is the number of degree of freedom.

5. The value ofX2 lies between zero and infinity, i.e., 0 < X2 < oo

6. The sum of two X distributions is again aX2 distribution, i.e., if~ and X2 are two independent X2 distributions, they have a X2 distribution with Y11 and 122 degrees of freedom respectively, then %\ + %\ is also a X2 distribution with (nl + n2> degree of freedom.

7. For different degrees of freedom, the shape of curve will be different.

8. Chi-square (X2> is a statistic hypothesis and not a parameter.

Working rule for Chi-Square Test

The chi-square test was first used by Karl Pearson in 1900 for testing statistical hypothesis. The computation is carried out in following steps in a sequential manner:

Step 1. Identification of problem.

Step 2. Data is arranged in the form of contingency table.

Step 3. Setting up of Null hypothesis (Ho) : According to null hypothesis, no association exists between attributes. This needs setting up of alternative hypothesis (HA). It asumes that an association exists between the attributes.

Step 4. Calculation of all, expected frequencies E corresponding to each cell of the contingency table, using the following formula:

Here n = Total sample size

C, = Sum total of column in which Ep lies

R, = Sum total of row in which Ep lies

Step 5. Take the difference between observed frequency 0 and corresponding expected frequency

E for each value of /, i.e. (0-E.).

Step 6. Square the values of (0-E) i.e. (0,- E;)2 for values of /= 1, 2, 3,... n.

Step 7. Divide each squared value by the corresponding expected frequency, i.e. calculate $\frac{O2}{E} - N$ for the values of/= 1, 2, 3,...n.

Step 8. Prepare a contingency table and enter frequencies, frequency differences, their squares and rvalues in the contingency table as follows:

Step 8. Add all the values obtained from Step 7. This represents the observed or calculated value of chi-square.

Step 9. Calculated X2 value is compared with tabulated value of~ at desired degree of freedom under different probabilities 0.5, 0.1, .05, .01, .001, etc.

Step 10. Inference : Conclusion is based on the following correlations :

1. rvalue is always positive because each difference is squared.

2. X2 will be zero, if each pair is zero. But it may assume any value from zero to infinity (0 to °°).

3. X2 is a static and not a parameter. It does not involve any assumption about the form of original distribution from which the observation comes.

4. Significance test on X2 is always based on one Hailed test on the right hand side of standard normal curve.

Method to draw Inferences:

1. If the calculated value of X2 is less than the tabulated value of X2, then the difference between observed and expected values is considered insignificant (appendix 4). But if the calculated value of X2 is more than the tabulated value then the two variables are dependent o" p8Ch other and value is significant.

2. The quantity in the denominator which is one less than the independent number of observations in a sample is called degree of freedom. If there are 2 classes (for example control and T, injected male and female) the degree of freedom would be 2 -l= 1. If there are three classes then d.f= 3-l= 2, in case of 4 classes d.f. = 4-I = 3 and so on.

3. If the X2 value is obtained in more than two pairs of data then d.f= (2 -l)x (2 -l)= 1.

2 x 2 Contingency Table

When data are cross-classified in such a manner that there are only two categories or two levels. These data are arranged in a table consisting of two rows and two columns. Such a table with two rows and two columns is called 2 x 2 contingency table.

In the following 2 x 2 contingency table the data is arranged in two rows and two

The formula used in chi-square analysis is :

$$X^2 = \Sigma \frac{(O-e)^2}{e}$$

where,

O is the observed value for a given category,

e is the expected value for the same,

Σ(sigma) represents the sum of calculated values for each category of the ratio,

O-e is the deviation in each case and can be represented by d.

Therefore, the equation can be reduced to :

$$X^2 = \Sigma \frac{d_2}{e}$$

For determining the results of inheritance of a particular character two things are to be emphasized:

(a) The experiment should include large number of individuals or plants.

(b) The experiment should be repeated for a number of generations and data should be properly maintained.

As a Test of Goodness of Fit

Chi-square test is applied as a test of "goodness of fit". Goodness fit indicates the closeness of observed frequency with that of the expected frequency. If the curves of these two distributions do not coincide or appear to diverge much, it is said that the fit is poor. If two curves do not diverge much, the fit is less poor. Thus it helps to answer whether something (physical or chemical factors) did or did not have an effect. If observed and expected frequencies are in complete agreement with each other then the chi-square value will be zero. But it rarely happens in biological experiments. There is always some degree of deviation.

Model problem :

In a Clinical treatment, the patients were tested to see the effect of a potential hypertensive drug. The 50 patients were assigned to receive active drug and other 50 as placebo at random. Their response to treatment was categorized as favorable or unfavorable. The data is given in the table below :

Table 1 Result of the effect of hypertensive drug.

Treatment	Response		Total
	Unfavourable	Favourable	
Placebo	41	9	50
Drug	18	34	50
	57	43	100

Test the hypothesis that drug has a significant effect. Use at $p = 0.05$

Solution:

Step 1. Problem identification : The hypertenisve drug has a significant effect

or not.

Step 2. Data: Given in the table. The attributes are arranged in two way table

or contingency table i.e. two rows and two columns.

Step 3. Hypothesis

1. Null hypothesis (Ho) stands for that drug does not have significant effect.

2. Alternate hypothesis (HA) proposes that the effect of drug is significant

Step 4. Level of significance =0.05 and Degree of freedom = (2-1) x (2-1) = 1

Step 5. Calculation of expected frequency for each by using following formula:

$$\text{Expected frequency } E = \frac{\text{Row total x Column total}}{\text{Grand total}}$$

$$\chi 2 = \sum \frac{(O-E)-0.5}{E} = 23.50$$

Inference: The calculated or observed value of χ^2 is 23.50. It is much higher than the critical value of χ^2 Therefore, the null hypothesis is rejected. The conclusion is that the hyper drug has significant effect.

10.1.2 Student t'test or t-test:

An Irish statistician, W.S. Gossett in 1908, applied this test for testing the significance of difference between the means of two different samples of small size. It was named 't-test'. It was further elaborated and explained in various ways by R.A. Fisher. Since, Gossett used to write under pseudonym 'student', t-test was named student t-test by William Seeley. Student Mest is also described as t-distribution or t-ratio. Student's t-test is applied to small samples :

Students t-test for a Single Population

Student r-test represents the ratio of the difference of population parameter and corresponding statistic and the standard error of the statistic with n-l degrees of freedom if the sample size is n. This can be represented as follows with (n-l) degrees of freedom if the sample size is n. The degree of freedom for t = size of sample -1. If the size of sample is represented by n then the degree of freedom will be n-l.

Difference of population parameter and corresponding statistic

Standard error of statistic

If a sample of size n is drawn from a normal population with mean (J, and variance 2 the student t will be:

$$t = \frac{\bar{X} - \mu}{\text{S.E. of Means}} = \frac{\bar{X} - \mu}{(\text{S.E.}_{\cdot M})\sqrt{s^2/n-1}}$$

Where

X = sample mean
// = population mean

n = sample size and

S = standard deviation of the sample

Student's t-test for two samples
The student t-test for two samples or for two variables is the ratio of difference between the means of samples and the standard error of difference between two means. The t-ratio is obtained by using the following formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{S.E}_D} \quad \text{or} \quad \frac{\text{Difference of two means}}{\text{Standard error of difference between two means}}$$

Mean of one variable or one sample
Mean of second variable or second sample

Standard error of difference between two means

Assumptions for t-Test
The t-test is applied under the following assumptions:

I •Samples are drawn from normal populations and are at random.

2. For testing the equality of two population means, the population variances are regarded as equal.

3. The population's standard deviation may not be known.

4. In case of two samples some adjustments in degrees of freedom for tare made. .

Properties of t-Distribution

1. t-distribution is a symmetrical distribution with mean as zero.

2. The graph off-distribution is similar to normal distribution except for the following two differences:

(a) Thenormaldistributioncurveishigherinthemiddlethanf-distributioncurve.

(b) f-distributionhas agreater spread sideways than the normal distribution curve, ltmeans thereismorearain thetai of t-istibution.

3. The f-distribution curve (Fig.33) is asymptotic to X-axis ,i.e. it extends to infinity on either side.

4. The shape of t-distribution curve varies with the degree of freedom.

5. The larger is the number of degree of freedom the more closely t-distribution resembles standard normal distribution.

6. Sampling distribution of r does not depend on population parameter. It depends on degree of freedom (n-1).

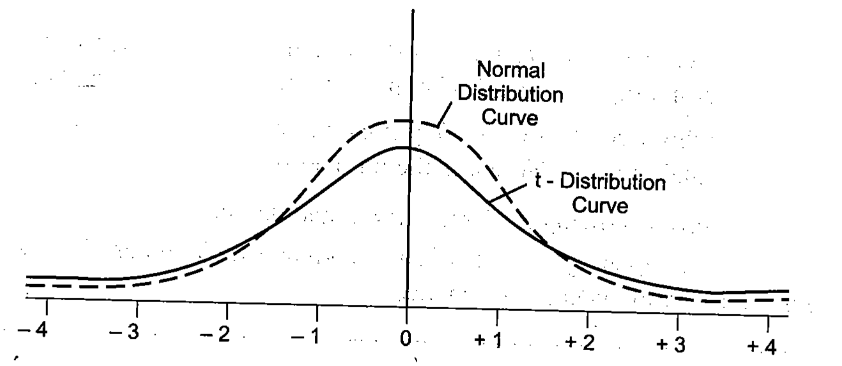


Fig 33: Diagram to show difference in the normal distribution curve and f-distribution

Applications of t-Distribution
The t-distribution has following three important applications in testing the hypothesis for small sample:

1. To test the significance of difference between two (5) sample means, the population variances being equal and unknown.

2. To test the significance of a single mean, when population variance o is unknown.

3. To test the significance of an observed sample correlation coefficient or difference between means of two samples (dependent sample or pairel observation).

10.1.2.1 Types of t-Tests:

Following two types of t-tests can be performed :

1. t-Test for Single Mean or t-Test for Independent Samples: The observations are classified nto groups. A test of mean difference is performed for a specified variable inis is called t-test for single mean.

2. t.Test for Paired Samples or t.Test for Two Sample Means: When paired observations are arranged case wise, and a test of treatment effect is performed, this is called two sample means or paired sample means. It is also known as 'correlated t-test'.

1. t-Test for Single Mean

If we assume that:

1. In a normal population of n size, the random samples collected are X X
2.SX=Xi+Xg+...X,,. "

3. Mean of all the samples X, + Xg + . ..X= x.

4. Value of hypothetical mean or population mean = mo.

5. Standard deviation =s.

6. Total number of values or size of sample = n

7. Degree of freedom =n -1

Thus in the above case :
1.Thevalueof-t'isdeterminedwhenwewishtocomparethemeanofasampledatawith agiven

specifiedvalue(iLi,sayheartbeatofman).

2. The critical value -f is determined at 5% level from the 'f distribution table at a suitable degree of freedom. In this case degree of freedom is n-1.

A.population of cats is known to have 160 heart beats per minute When 13
cats were each fed on a fixed quantity of a drug and data taken on their beats, When 3 x = 147 with S = 27.5. Find if there is a change in heart beat due to drug.

Data Given:

Sample mean x =147

Population mean //o = 160

StandarddeviationS=27.5

Sample size n =13

Degree of freedom (n-1) = 13-1= 12

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{147 - 160}{27.5/\sqrt{13-1}}$$

$$= \frac{13}{27.5/3.46} = \frac{13}{7.94} = \mathbf{1.64}$$

Step 1. From the table of two tailed test t at0.05,/.e., at of degree of freedom fi2 = 2.179.
Step2.Foronetaiedtestt,, =1.782at0.05/.e.at10degreeoffredom.

Step3. rvalue for two tailed test is used only if null hypothesis (Ho) is//= r110 = 160.

Step 4. rvalue for one tailed test is used only when n < Up or null hypothesis is not applicable i.e.

Step 5. Critical ratio : In this sample

(1) the estimated or calculated value of fi2 1.64.

(2) The hypothetical value of fi2= 2.179.

Inference : The estimated value of t is less than the hypothetical value of t but the difference is not significant. Therefore, Null hypothesis (Hp): jLl = I-lois accepted because enough reason to reject it is not available. It means the heart beat rate due to drug did not change significantly.

To determine the average height of the plants, a random sample of 25 plants

was taken. The following results are available :

X =63"

S =4"

Solution:

Can we assume that the average height of plants is 60"?

$$= \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{(63-60)}{4/\sqrt{25}} = \frac{3}{4/5}$$

$$= \frac{3 \times 5}{4} = 3.75$$

The hypothetical rvalue from the Hable at 5% significance level and 24° of freedom = 2.797.

The Critical Value:

1. Hypotheticalvalueoffat5%significancelevel=2.797.

2. Calculated rvalue at the same level = 3.75.

Inference : The calculated value of 't' is more than the hypothetical 't' value given in the table. The difference is significant. Thus, the average length of the trees cannot be 63". Therefore, Null hypothesis is not applicable.

2. V Test for Two Sample Means or Correlated t-test or Paired t-test Paired t-test is applied to paired data or two samples obtained from the same population at two different times and conditions. Each individual gives a pair of observations. The Mest for correlated data provides the significance of the difference between two correlated means. Correlated Mests are mostly applicable to biological experiments and field studies.

To calculate significance of difference between two correlated means, the following formula is used:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE_D}$$

$$S.E_D = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

0

$$S.E_D = \sqrt{SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2} - 2 \times r \times SE_{M_2}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S\sqrt{\frac{1}{n_2} + \frac{1}{n_2}}}$$

where, x,xg= Means of sample data
SED= Standard error of difference

SD= Standard deviations of two samples

n-1=Total number of values of two samples

In the above case degree of freedom is taken as: .

df={n- I)+(n-1) or (n+n)- 2

SED, i.e.. Standard Error of difference is calculated by using following formula

Here, SE, is (SE) for first test mean.

SE, is standard error (SE) for second test mean.
and r is correlation coefficient between scores made on first and second tests.

The percentage of water content in two varieties of water melons was measured and the followina results were obtained :

Find out whether there is significant difference in water content of two varieties. Solution:

Identification of Problem: To find out difference in the water content of two varieties of water melons.

Data Given :

$n_1 = 12$	$n_2 = 15$
$S_1 = 15$	$S_2 = 19$
$\bar{X}_1 = 92$	$\bar{X}_2 = 84$

Step 1 : Calculation of standard deviation:

$$S = \frac{(n_1 - 1)\cdot S_1^2 + (n_2 - 1)\cdot S_2^2}{n_1 + n_2 - 2} = \frac{(12-1)\cdot(15)^2 + (15-1)\cdot(19)^2}{12 + 15 - 2}$$

$$= \frac{2475 + 5054}{25} = \frac{7529}{25} = 301.16$$

Step 2 : Calculation of t-value :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{92 - 84}{\sqrt{\frac{225}{12} + \frac{361}{15}}}$$

$$= \frac{8}{\sqrt{18.75 + 24.06}} = \frac{8}{\sqrt{42.81}} = \frac{8}{6.54} = 1.22$$

Step 3 : Critical Value :

1. Hypothetical value of from distribution table = 1.708.

2. Calculated value of from observations = 1.19.

Inference : Observation of estimated t value 1.19 is less than hypothetical value of t which is 1.708. Therefore, it can be concluded that there is no significant difference in water contents in two varieties of water melons. Hence, Null hypothesis is applicable or true for this case.

t-test from Grouped Data

For -test from grouped data, the raw data is arranged into groups in ascendina series The Sam formula is used except for following modified formula

$$\bar{X} = \frac{\Sigma f \cdot x}{\Sigma f} \text{ and } \sigma x = \sqrt{\frac{\Sigma f - x^2 \frac{(\Sigma f \cdot x)^2}{\Sigma f}}{\Sigma f - 1}}$$

To find out the effect of a hormone spray on the yield of trench beans, following results were obtained. Draw inference using 'f test as regards to the hormonal spray on the seed yield.

X_1(control) : 30, 35, 31,36, 38, 32, 25, 39, 31, 34, 33, 35, 40, 30, 32, 28, 26, 29, 30, 30, 35, 36, 37, 38, 39, 30, 31, 39, 40, 35, 36, 26, 27, 40, 41, 35, 38, 32, 31, 38, 30, 35, 36, 25, 28, 28, 27, 36, 36, 30.

X_2 (treated) : 35, 38, 40, 45, 40, 42, 54, 55, 43, 45, 50, 44, 51, 52, 53, 52, 48, 50, 49, 47, 49, 50, 49, 51, 52, 50, 51, 52, 40, 46, 48, 49, 50, 35, 38, 45, 47, 45, 46, 50, 51, 46, 41, 42, 48, 49, 51, 52, 50, 53.

Identification of Problem : To see effect of hormone spray on the seed field in trench beans.

(1) Null hypothesis : Effect of hormone spray is insignificant. Therefore,
$Ho=\mu=\mu_1$

(2) Alternative hypothesis : Effect of hormone spray is significant and Null hypothesis is not accepted

Solution: Therefore, $H_1=\mu=\mu_1$

Step 1. Data Given

The values (number of seeds/plant) are arranged in ascending order of magnitude with their frequency values.

Table 6.6 Number of seeds on treated and untreated plants

Solution :

1. Total no. of control plants = $\Sigma f_1 = 50$

2. Total no. of treated plants = $\Sigma f_2 = 50$

3. Total no. of seeds on control plants = $\Sigma(f_1 X_1) = 1670$

4. Total no. of seeds on treated plants = $\Sigma(f_2.X_2) = 2359$

Step 2. Calculation of mean of X_1 and X_2:

$$\bar{x}_1 = \frac{\Sigma(f_1 \cdot x_1)}{\Sigma f_1} = \frac{1670}{50} = 33.4$$

$$\bar{x}_2 = \frac{\Sigma(f_2 \cdot x_2)}{\Sigma f_2} = \frac{2359}{50} = 47.18$$

Step 3. Standard deviation of X_1 and X_2:

$$\text{SD or S of } x_1 = \sigma x_1 = \sqrt{\frac{\Sigma f_1 \cdot x_1^2 - \left(\frac{\Sigma f_1 \cdot x_1}{\Sigma f_1}\right)^2}{\Sigma f_1 - 1}} \quad \text{or} \quad \sigma^2 x_1 = \frac{\Sigma f_1 \cdot X_1^2 - \frac{\Sigma (f_1 \cdot x_1)^2}{\Sigma f_1}}{\Sigma f_1 - 1}$$

$$= \frac{56726 - \frac{(1670)^2}{50}}{50 - 1}$$

$$= \frac{56726 - 55778}{49} \quad \text{or}$$

$$\sigma^2 x_1 = \frac{984}{49} = 19.347$$

$$\text{Similarly,} \quad \sigma x_2 = \sqrt{\frac{\Sigma f_2 \cdot x_2^2 - \frac{(\Sigma f_2 \cdot x_2)^2}{\Sigma f_2}}{\Sigma f_2 - 1}} \quad \text{or} \quad \sigma^2 x_2 = \frac{\Sigma f_2 \cdot x_2^2 - \frac{(\Sigma f_2 \cdot x_2)^2}{\Sigma f_2}}{\Sigma f_2 - 1}$$

$$\text{or} \quad \sigma^2 x_2 = \frac{112499 - \frac{(2359)^2}{50}}{50 - 1}$$

$$= \frac{112499 - 11297.62}{49}$$

$$\sigma^2 x_2 = \frac{1201.38}{49} = 24.518.$$

Step 4. Combined variance^2 of X_1 and X_2:

$$\sigma x_1 + x_2 \quad \text{or} \quad \sigma_d = \frac{19.347}{50} + \frac{24.518}{50}$$

$$= 0.387 + 0.49 = 0.877.$$

Step 5. Combined standard error of difference ($S.E._D$):

$$S.E._D \text{ or } \sigma d = \sqrt{\sigma_d^2} = \sqrt{0.877} = 0.936$$

$$\bar{x}_1 - \bar{x}_2 = 33.4 - 47.28 = 13.88$$

Step 6. Value of t :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_d} = \frac{13.88}{0.936} = 14.829.$$

$$\text{d.f.} = N_1 + N_2 - 2$$
$$= 50 + 50 - 2$$
$$= 98$$

Calculated value off comes to = 14.829.
Table value of ton dfgs at 5.0% level = 1.96.

Inference: The calculated value of 'f 14.829 is very high than the tabulated't value 1.96. Therefore, we can say that the hormone spray definitely has significant effect on the seed yield.

Exercises

1. The heights of ten children selected at random from a given locality had a mean 63.2 cm and variance 6.25 cm. Test at 5% level of significance, the hypothesis that the children of the given locality are on the average less then 65 cm in all. Given for 9 degrees of freedom P(t> 1.83) = 0.05.

Hint: Alternative hypothesis H, is accepted.

2. Body length of 10 fishes of a species of fish population was recorded from two ponds in the following table:

Pond	Length of fishes in cm									
A	20	24	20	28	22	20	24	32	24	26
B	12	10	8	10	6	4	14	20	10	6

Calculate the mean difference in total body length between the fishes in the above two ponds is significant or not. . ':

Hint: Calculated value of tcomes o 7.337, and hypothetical value oft atd.f. 18 on 0.1 level is 2.552. The calculated value is more than hypothetical value. Null hypothesis is not applicable.

3. In a pig farm a group of 12 pigs were fed with one variety of food A while another group of 12 pig: of same age, sex and stock were given another variety of food B. After one month weights of both

First group on food A	**31**	**34**	**34**	**29**	**26**	**32**	**35**	**38**	**40**	
First group on food B	**26**	**24**	**28**	**29**	**30**	**29**	**32**	**26**	**35**	**29**

Whether the difference in mean weights of first and second group (due to different food quality) is significant or not. For your answer use students t' test.

Hint: Difference is significant.

10.1.3 ANOVA : (Analysis of Variance)

Test of significance or 't test' is used to compare the means of two samples. The differences in the means of more than two samples or populations can not be tested by these methods or by methods discussed in the previous chapter. The statistical technique used to compare means of variations of more than two populations is called 'analysis of variance (ANOVA). The concept of ANOVA or analysis of variance was introduced by R. A. Fisher. ANOVA is based on two types of variations:

1. Variations existing within the sample

2. Variations existing between the samples.

The ratio of these two variations is an indication of sample differences. Therefore, ANOVA helps in estimating whether more variations exist among the group means or within the groups. The ratio of these two variations is denoted by 'F'. The F- value is the indication of differences between the samples.

ANOVA is based on following assumptions :

1. All ANOVA require random sampling.

2. The items for analysis of variance should be independent. But this may not be true in case there is correlation in time or space.

3. Equality of variances in a group of samples is an important precondition for several statistical tests.

4. The residual components should be normally distributed.

5. The variable of interest for each population or group has a normal probability distribution.

6. The variance associated with the variable of interest must be the same for each population or for each group of data.

Steps involved in ANOVA:

For performing the test of analysis of variance following statistical steps are carried out :

1. Sum of Squares (S.S.)

Sum of squares is computed as the sum of squares of deviation of the values for mean of the

sample. It means:

$$S.S = \Sigma\ (X\text{-}X)^2 \text{ or}$$

$$S.S = \Sigma x^2 - \frac{\Sigma x2}{n}$$

Here, X represents sum of squares

X represents mean of sample

X represents deviation

n represents total number of observations.

2. Mean Square (M.S.)

Mean square is the mean of all the sum of squares. It is obtained by dividing S.S. by appropriate degree of freedom (df):

$$M.S = \frac{\text{Sum of squares}}{\text{Degree of freedom}} = \frac{S.S.}{df}$$

Therefore, following three values or quantities are required for computing ANOVA :

1. Total sum of squares
2. Between sample sum of squares
3. Residual sum of squares.

10.1.3.1 One Way Analysis of Variance:

Simplest type of analysis of variance is known as one way analysis of variance. In one way ANOVA only one source of variation (or factor) is investigated. But investigations are carried out in three or more samples simultaneously.

One way analysis of variance is used to test the null hypothesis that three or more treatments are equally effective.

Table 5 Hemoglobin levels (%) of children fed with three different diets

S.NO.	Group I	Group II	Group III
1	11.6	11.2	9.8
2	10.3	8.9	9.7
3	10.0	9.2	11.5
4	11.5	8.8	11.6
5	11.8	8.4	10.8
6	11.8	9.1	9.1
7	12.1	6.3	10.5

8	10.8	9.3	10.0
9	11.9	7.8	12.4
10	10.7	8.8	10.7
11	11.5	10.0	-
12	-	9.7	-
	t_1 = 124.0	t_2 = 107.5	t_3 = 106.1

Test whether the means of these three groups differ significantly.
Solution:

No. of observations	$n_1 = 11$	$n_2 = 12$	$n_3 = 10$
Total of all samples	$t_1 = 124.0$	$t_2 = 107.5$	$t_3 = 106.1$
Sample Mean	$\bar{X}_1 = \frac{124.0}{11} = 11.27$	$\bar{X}_2 = \frac{107.5}{12} = 8.96$	$\bar{X}_3 = \frac{106.1}{10} = 10.61$
	$S_1^2 = 124 \times 11.27 = 1397.48$	$S_2 = 107.5 \times 8.96 = 963.2$	$S_3 = 106.1 \times 10.61 = 1125.72$

$$N = n_1 + n_2 + n_3 = 11 + 12 + 10 = 33$$

$$\Sigma X \quad \text{or} \quad T = t_1 + t_2 + t_3 = 124.0 + 107.5 + 106.1 = 337.6$$

Common Mean $$\bar{\bar{X}} = \frac{124.0 + 107.5 + 106.1}{33} = \frac{337.6}{33} = 10.23$$

The common mean x of all the samples is 10.23 whereas mean of sample T, (/.e. X,) is 11.27, of sample Tg (/.e. Xg) is 8.96 and mean of sample Tg (i-e- Xs) is 10.61. We have to see whether the mean levels of 3 different samples in comparison to common mean x show significant variance.

Null Hypothesis

Step 1. Calculation of three sums of squares or within group sum of squares:

1. Sum of observations from all 33 individuals from all three samples:

= (ΣX) = 11.6+10.3+10+...

12.4+10.7 = 337.6

2. Sum of squares of all 33 observations =

$$\Sigma X^2 = (11.6)^2 + (10.3)^2 + (10.0)^2 \ldots$$

$$(12.4)^2 + (10.7)^2 = 3516.32$$

3. Correction term = $\frac{(\Sigma X)^2}{n} = \frac{(337.6)^2}{33} = \frac{113973.76}{33} = 3453.75$

4. Total sum of squares = $S(X - \bar{X}^2)$

$$= SX^2 - \frac{(\Sigma X)^2}{n}$$

$$= 3516.32 - 3453.75 = 62.57$$

Step 2. Calculation of sum of squares between groups, i.e., among groups sum of squares: It is the relation between the common mean and the mean for each group, with due allowance for the number of observations in each group. Square the total for each group and divide each square by the number

of observations in that group. Total the quotients and subtract the correction term. Sum of squares between groups is =

$$\frac{(124.)2}{11} + \frac{(107.5)2}{12} + \frac{(106.1)2}{10} - 3453.75$$

=32.81

Step 3. Sum of squares within groups is found by subtracting item 2 from item 1, i.e., sum of squares within the group = Total sum of squares - between group sum of squares = 62.57-32.81 = 29.76

Degrees of Freedom:

We know that the degrees of freedom are one less than the number of items and the sum of squares.

1. Total sum of squares = No. of observations = 33 -1

The degree of freedom (d.f.) = 33 -1 = 32

2. Sum of squares between groups = No. of groups = 3

The degree of freedom (d.f.) = 3-1 = 2

3. Sum of squares within groups = Total sum of squares - Sum of squares between groups = 32 - 2 = 30

4. Mean square = $\frac{\text{Sum of squares between the groups}}{\text{Degree of freedom}}$

(a) Variance (sum of squares) between the groups = $\frac{32.8}{2} = 16.405$

(b) Variance (sum of squares) within the groups = $\frac{29.76}{30} = 0.992$

The results are represented in the following table :

Table 2 Table for the Analysis of Variance

$$\text{Variance Ratio, Fobs} = \frac{\text{Mean square between groups (M.S.)}}{\text{Mean square within groups (Residual)}}$$

$$\frac{16.405}{0.992} = 16.54.$$

Inference

From the above results, it is evident that between the groups, variability is greater than the within group variability. Also all the three groups are not from populations having identical means.

The F-Test

F test is comparison of the means of samples from two or more populations. It is comparison of the mean square between samples and mean square within samples :

$$F = \frac{\text{Mean sauare between samoles}}{\text{Mean sauare within samoles}}$$

The sampling distribution of S2, C2 is called F distribution. It is determined by comparing observed value ofF statistics to the critical value of F at 5% level from the F distribution table. When null hypothesis is true or all the means of different populations are equal, SJ and S 2 are both unbiased estimates of cr~ (variance), we can expect the ratio S~/szw to be near 1. When the null hypothesis is not true, it tends to be larger and over-estimates μ. In such cases S^2 /sZ, will tend to be larger. It means a high value of F value indicates the significant differences between the samples. Therefore, F represents variance ratio. The F-test was first developed by R. A. Fisher. The F-distribution is used to test the hypothesis about the variance of two populations. The analysis of variance technique is used to find how much of the variation in the observations is due to between sample differences and how much is due to random variability within the samples. By comparing these two variations we can determine the importance of the sample differences. The same analysis of variance technique can also be used when more than one factors affect the responses.

Assumptions in F-Test

1. The values in each group should be normally distributed.

2. The variances within each group should be equal for all groups.

3. The variances of each value around its own group mean, (i.e., error) should be independent for each value.

EXERCISES

1. Fishes were reared in three different ponds with different types of food. A sample of 5 fishes was selected from each pond. Their weights are recorded in the table below. Find out if these data suggest a difference in average weight of fishes reared indifferent ponds.

2. The three varieties of wheat A, B and C were sown in four agricultural plots. Their yield in quintals per acre is recorded in the table below. Test the significance of difference between the yields.

3. The following table shows the emotional maturity scores of 27 young adult males classified by aqe and the use of marijuana.

Perform an analysis of variance of these data.

10.1.4. Correlation and Regression Analysis :

Various statistical methods studied so far, like measures of central tendency, average, measures of dispersion and skewness are related to one variable only. There are many situations where two variables are inter-related and a change in the value of one variable causes change in the value of other variable. For example, we may like to study the relationship between height and weight of persons, blood pressure and age, consumption of certain nutrient and weight gain or intensity of stimulus and reaction time or intensity of stimulus and intensity of reaction. The study of nature and strength of relationship between two variables is described in terms of correlation and regression.

In correlation analysis, we are concerned whether two variables are independent or they vary together in positive or negative direction. In correlation the two variables are not related as independent and dependent variables. It means in correlation both the variables are affected by a common cause and the degree to which these variables vary together is estimated. In regression analysis, the dependence of one variable on another variable is determined. Therefore, the two variables are related as independent and dependent variables. Regression analysis is employed to predict or estimate the value of one variable corresponding to a given value of another variable. Regression equations are applied to determine changes in Y due to changes in X variable.

10.1.4.1 Correlation:

Correlation is the tendency of simultaneous variation between two variables'. According to Connor 'if two or more quantities vary in sympathy and the movements in one tend to be accompanied by corresponding movements in the other, then these two quantities are said to be correlated.

Explanation

1. Correlation indicates the degree of relationship between two variables.

2. The movements in one variable are accompanied by corresponding movements in the other variable.

3. According to Tuttle, correlation is an analysis of co variation between two or more variables.

Significance of Correlation

The study of correlation is of great significance in practical life, because of the following reasons:

1. The study of correlation enables us to know the nature, direction and degree of relationship between two or more variables.

2. Correlation studies help us to estimate the changes in the value of one variable as a result of change in the value of related variable. This is called regression analysis.

3. Correlation analysis helps us in understanding the behavior of certain events under specific circumstances. For example, we can identify the factors for rainfall in a given area and how these factors influence paddy production.

4. Correlation facilitates the decision making in the business world. It reduces element of uncertainty in decision-making.

5. It helps in making predictions.

10.1.4.2 Types of Correlation:

Depending on its extent and direction the correlation between two variables may be of following types. The positive and negative correlations are based on the direction of change in the value of two variables:

Perfect Positive Correlation : When two variables move proportionately in the same direction, i.e., the increase in the values of one variable leads to corresponding increase in the values of other variable, the correlation between them is called perfect positive. For example, increase in body weight with the increase in height presents positive correlation. It is also called direct correlation. Examples of perfect positive correlation are very rare in nature but some examples are:

- Correlation in day length and temperature, and
- Correlation in rain and humidity, etc.

Moderately Positive Correlation: When two variables are partially positively correlated, the correlation is termed moderately positive correlation, e.g., tallness of plants and the quantity of manure used, nutrition and death rate in pregnancy, and the mortality rate and overcrowding, etc.

Perfect Negative Correlation: The two variables show negative correlation when one variable increases with a constant interval and another decreases with constant interval. Thus,. variables deviate in opposite directions. This is also called inverse correlation. Examples of perfect

negative correlation are very rare in nature but some approaching to that extent are temperature and lipid content of the body, number of red blood corpuscles, Hb percentage, etc (Fig.34).

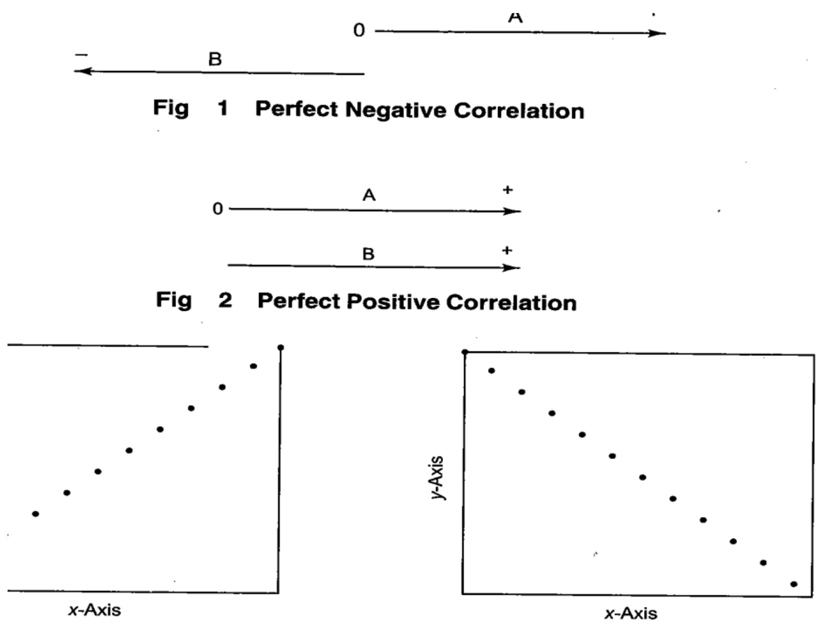


Fig.34 Perfect positive association(r=+1) & Perfect negative association(r =-1).

Following data presents perfect negative correlation :

Linear and Non-linear Correlations

The correlation can also be classified as linear or nonlinear on the basis of ratio of variations in the related variables:

Linear Correlation **:** Correlation between two variables is said to be linear if there is some constant relationship between the two variables. When the values of two variables are plotted as points in the XY plane, a straight line is formed. Linear correlation is very rare in biological observations

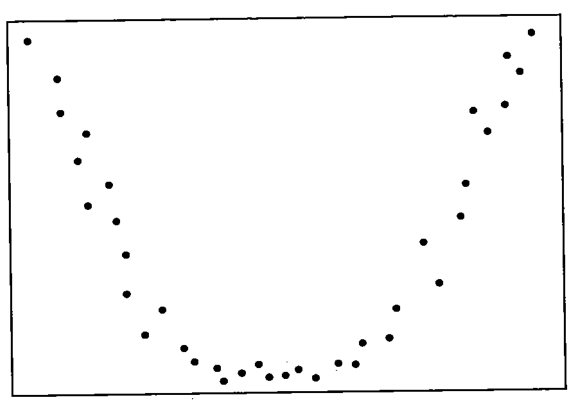


Fig. 35. Nonlinear association.

Non-linear Correlation **:** The relationship between two variables is said to be nonlinear or curvilinear if corresponding to a unit change in one variable, the other variable does not change at the same constant rate but fluctuates. It means ratio of variations in the values of the two variables is not constant (Fig.35)

Simple, Partial and Multiple Correlation

Based on the number of variables involved, the correlation may be of following three types :

Simple Correlation : In simple correlation only two variables are involved. Therefore, in simple correlation the relationship is between two variables such as intelligence of students and their performance (marks) in the examination.

Multiple Correlation: In multiple correlation relationship between three or more variables is studied. Simultaneous study of relationship between yield of wheat per acre, the amount of rainfall and the amount of fertilizer applied are the examples of multiple correlation.

Partial Correlation: In partial correlation, relation between more than two variables is considered but correlation is studied only between two variables. Other variables are assumed to be constant. For example, the correlation between the amount of fertilizers and the yield of wheat per acre is partial correlation in case rainfall is assumed to be normal.

10.1.4.3 Measures of Correlation:

Correlation analysis measures the degree of association of two variables. Following methods are used to measure the correlation between two variables : (Fig.36)

1. Scatter diagram method

2. Karl Pearson's coefficient of correlation

3. Spearman's rank correlation coefficient

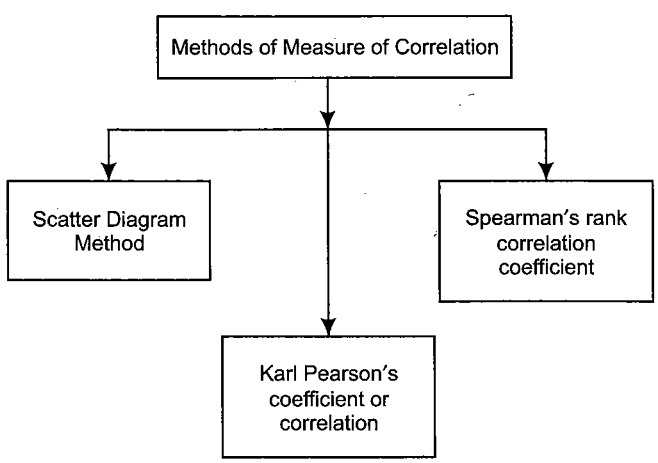


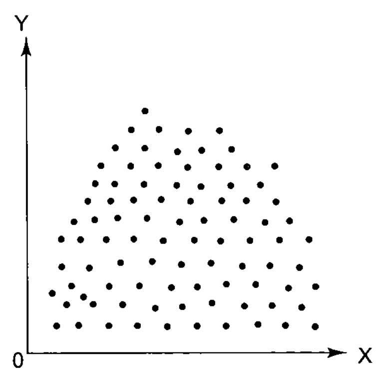
Fig.36 . Diagrammatic representation of different methods of measure of correlation.

1. Scatter diagram Method or Scatter Plot Method:

Scatter diagram is the simplest method of studying relationship between two variables. It is in the form of graphic representation of degree and direction of correlation between two variables.

Say we take two variables X and Y for n number of samples (X_1, X_2, X_3...... X_n) and plot X_1 against y_1 as a dot (.) in the XY-plane, the diagram of dots so obtained is known as scatter diagram or dot diagram. It is customary to take dependent variable along Y-axis, i.e., along vertical axis and independent variable along X-axis or horizontal axis. Placement of dots on the graph reveals whether the changes in the variable are in the same direction or in opposite direction.

The following scattered diagrams depict different forms of correlation:



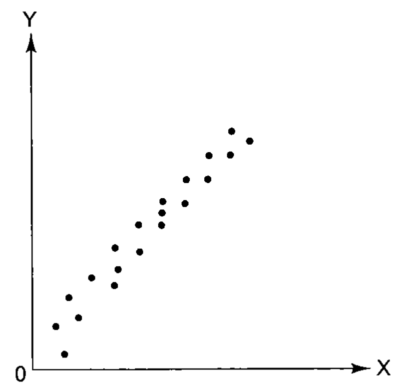
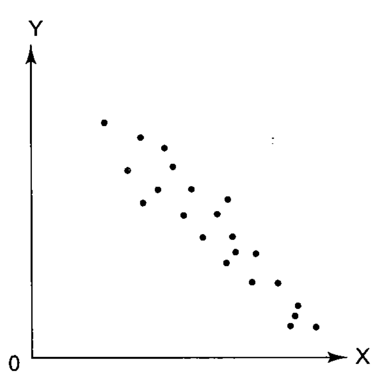


Fig.37 a: Scatter diagram with zero correlation Fig. 37 b:Scatter diagram with very +ve correlation.



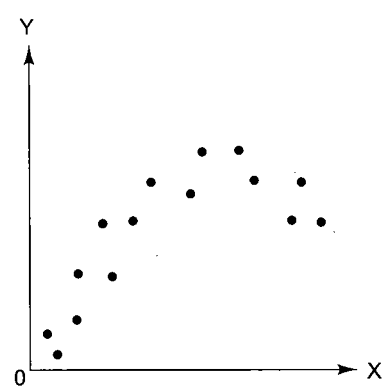


Fig.37 c Scatter diagram with very -ve correlation,

Fig.37 d Scatter diagram with curvilinear correlation.

2. Karl Pearson's Correlation Coefficient:

A scattered diagram, like a histogram, is a convenient way of displaying existence of correlation and direction of correlation, but it does not give any correlation value. To measure correlation, the coefficient of correlation is worked out by Karl Pearson's Coefficient of Correlation. Coefficient of correlation is the degree to which two variables are inter-related. It is a mathematical method for measuring the tendency of linear relationship between two variables. This was introduced by Karl Pearson (1867-1936). This measure of correlation is also known as Pearsonian Correlation Coefficient. If two variables are denoted by X and Y, the coefficient of correlation between them is represented by rxy or r. Pearson's coefficient correlation method can be used to measure correlation for individual series as well as for grouped data.

Properties of Coefficient of Correlation

1. Coefficient of correlation is a measure of closeness between two variables.

2. The correlation may be positive or negative.

3. The range of correlation coefficient is from -I to+1.

4. If r =+1, the correlation between two variables is perfect and positive.

5. If r= -I the correlation is perfect and negative(Fig.38a).

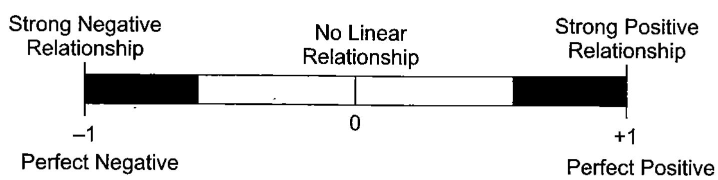


Fig.38a Coefficient of Correlation

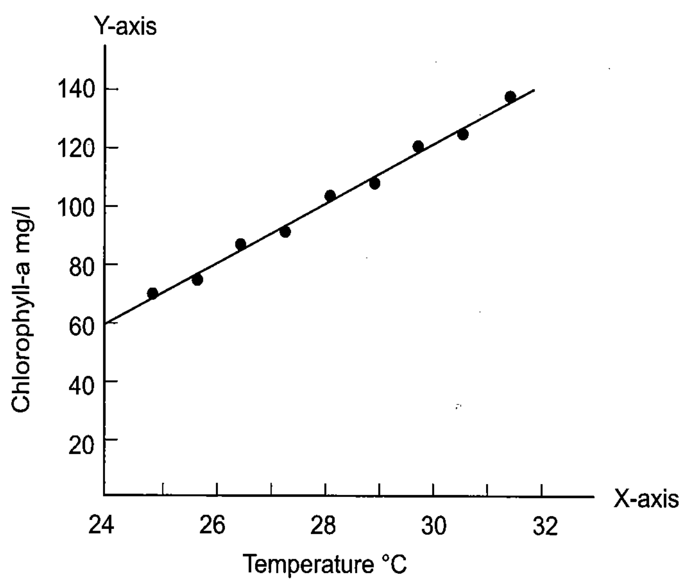


Fig. 38b Diagram to show range of correlation coefficient.

6. If there is a strong positive linear relationship between two variables, the value of r will be close to+1.

7. If there is strong negative linear relationship between the variables, the value of r will be close to-1.

8. If r = 0, there is no correlation between two variables. It means variables are independent.

The value of r is calculated by the following formula .

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}} \quad \text{or} \quad \frac{\Sigma x-\bar{x}\Sigma y}{\sqrt{(\Sigma x^2-\bar{x}\Sigma x)(\Sigma y^2-\bar{y}\Sigma y)}}$$

Computation of Coefficient of Correlation from Ungrouped Data:

Coefficient of correlation for unsrouned data is calciilatpri hv the fnllnu/ino formula

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}} \quad \text{or} \quad r = \frac{\Sigma\left(\frac{x-\bar{x}}{\sigma x}\right)\left(\frac{y-\bar{y}}{\sigma y}\right)}{n}$$

Here, x and y are measurements for two variables,
x and y are means of two distributions of measurements and

(σx and σy are standard deviations of measurements.

For five patients, temperature (J<) and pulse (y) are given. The correlation of

coefficient for these two measurements can be calculated as follows :

Patient	Temperature x	Pulse y	$\left[\frac{x-x}{\sigma x}\right]$	$\left[\frac{y-y}{\sigma y}\right]$	$\left[\frac{x-x}{\sigma x}\right]\left[\frac{y-y}{\sigma y}\right]$
A	**102**	**100**	$\frac{102-100}{\sqrt{2}}=\frac{2}{\sqrt{2}}$	$\frac{100-80}{\sqrt{200}}=\frac{20}{\sqrt{200}}$	$\frac{40}{\sqrt{400}}$
B	**101**	**90**	$\frac{101-100}{\sqrt{2}}=\frac{1}{\sqrt{2}}$	$\frac{90-80}{\sqrt{200}}=\frac{10}{\sqrt{200}}$	$\frac{10}{\sqrt{400}}$
C	**100**	**80**	$\frac{100-100}{\sqrt{2}}=\frac{0}{\sqrt{2}}$	$\frac{80-80}{\sqrt{200}}=\frac{0}{\sqrt{200}}$	$\frac{0}{\sqrt{400}}$

D **E**	**99** **98**	**70** **60**	$\frac{99-100}{\sqrt{2}} = \frac{-1}{\sqrt{2}}$ $\frac{98-100}{\sqrt{2}} = \frac{-2}{\sqrt{2}}$	$\frac{70-80}{\sqrt{200}} = \frac{-10}{\sqrt{200}}$ $\frac{60-80}{\sqrt{200}} = \frac{-2}{\sqrt{200}}$	$\frac{10}{\sqrt{400}}$ $\frac{40}{\sqrt{400}}$
Total	**Σx =500**	**Σy =400**	**0**	**0**	$\frac{40}{\sqrt{400}}$

$$\bar{x} = 500/5 = 100$$

$$\bar{y} = 400/5 = 80$$

$$s_x = \sqrt{2}$$

$$s_y = \sqrt{200}$$

$$r = \frac{\left[\frac{x-\bar{x}}{\sigma x}\right]\left[\frac{y-\bar{y}}{\sigma y}\right]}{n}$$

$$= \frac{100/\sqrt{400}}{5} = \frac{100/20}{5} = \frac{5}{5}$$

$$= 1$$

When coefficient of correlation's value is+1, it means there is perfect and positive correlation, i.e. the two variables, i.e., temperature and pulse rate change in the same direction.

10.1.5. Regression:

In analyzing data, we find that it is frequently desirable to learn something about the relationship between two variables. For example, we may be interested in studying the relationship between blood pressure and age, height and weight, the concentration of an injected drug and heart rate, the consumption level of some nutrient and weight gain, the intensity of a stimulus and reaction time or total family income and medical care expenditures. The nature of relationship between variables such as these may be examined by regression analysis (Fig.39).

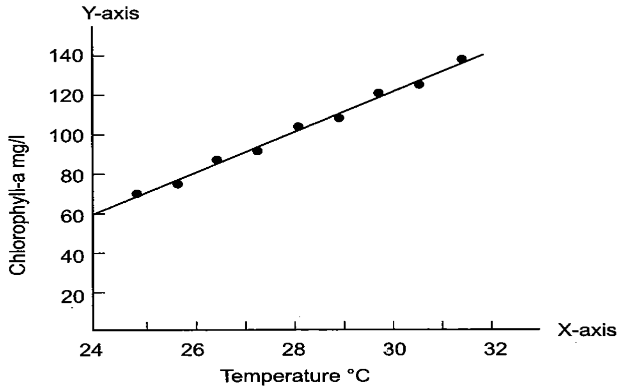


Fig.39 Regression analysis and correlation of the data.

The term regression was coined by F. Galton inl885 to explain the data obtained during the study of inheritance. Galton observed the height of offspring's during a few generations of a family and came to the conclusion that the height of Offspring tend to occupy median position. He expressed the regression as 'the tendency to remain towards central position.

10.1.5.1 Objectives of Regression Analysis:

I. The regression analysis is used to predict the value of one character or variable from the value of the other character or variable. According to this, the variables may be:

(a) Dependent Variable: The variable whose value is influenced or is to be predicted, is called a dependent variable.

(b) Independent variable; The variable which influences the values is called an independent variable.

2. The regression analysis is used to find out the measures of error present during the use of regression line for prediction. For this, standard error of estimate is calculated.
3. From the value of coefficient of correlation, one can find the degree of association between two variables, but from the regression analysis one can predict how a change in one variable is expected to affect the other.

10.1.5.2 Types of Regression Analysis:

The regression analysis can be of two types: simple and multiple.

1. Simple Regression: The regression analysis confined to the study of only two variables at a time is termed as simple regression.

2. Multiple Regression: The regression analysis for studying more than two variables at a time is known as multiple regression.

Regression Lines and Linear Regression

When observations from two variables are plotted as a graph, and if the points so obtained fall in a straight line then the relationship is linear and it is said that there is linear regression between the variables under study. However, if the line is not a straight line, the regression is termed as non- linear regression (Fig.40).

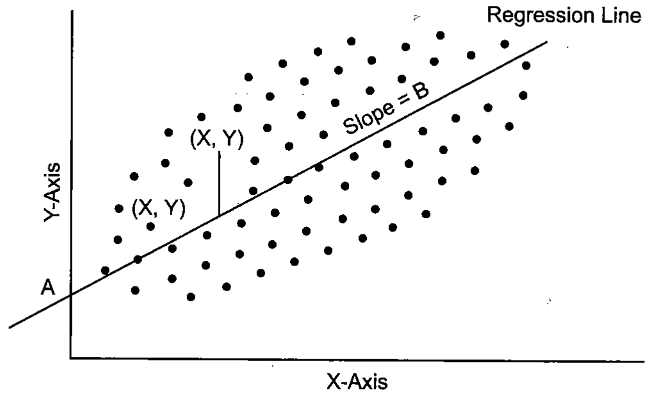


Fig.40 Regression line drawn with different values of X and Y.

When the points are obtained on a scattered diagram, the process of deciding the line of the best fit to summarize a particular set of points on a graph is called regression analysis. This is worked out by deriving an equation called regression equation.

Assumptions Underlying Simple Linear Regression

In the simple linear regression model the two variables are X and Y. The variable X is usually referred to as the independent variable. The variable Y is called the dependent variable. Their relation is called regression of Yon X. The assumptions of simple linear regression are :

1. The independent variable is also called nonrandom or mathematical variable.

2. Values of independent variable are said to be 'fixed'. It means the values of X are preselected.

3. Variable X is measured without error.

4. For each value of X there is a subpopulation of Y values.
5. The variances of the subpopulations of Y are all equal.

6. The means of subpopulations of Y all lie on the same straight line. This is called assumption of linearity.

7. Geometrically, a and prepresent the y-intercept and slope respectively on the line on which all means are assumed to lie.

8. Y-values are statistically independent. It means the values of Y chosen at one value of X do not depend on the values of Y chosen at another value of X.

10.1.5.3 Regression Equation :

The equation that describes position of any line on a graph is called regression equation. For a linear regression, the equation for a dependent variable XY against independent variable x can be given as follows:

$$y = a + bX$$

Here, values of 'a' and 'b' are constant and are fixed for a particular line. If the values of 'a' and 'b' are known, y can be obtained for any corresponding value of X. The values of 'a' and 'b' are cal cul ated hv the following equation :

$$b = \frac{\Sigma(X - \bar{X})(y - \bar{y})}{\Sigma(X - \bar{X})^2} = r\frac{(\text{SD of } y)}{(\text{SD of } x)} \quad \text{or} = r\frac{\sigma_x}{\sigma_y}$$

$$\text{or} \quad b = \frac{\Sigma xy - \bar{X}\,\Sigma y}{\Sigma X^2 - \bar{X}\,\Sigma X}$$

After obtaining the value of 'b' the value of 'a' can be calculated.
The constant 'a' is known as intercept, and denotes the value of y when the value of xis zero.

The constant 'b' measures the slope of the line and is called "regression coefficient". The constant 'b' gives an idea of that how change occurs in variable ywhen the variable X varies by 1 unit. For instance, if the value of 'b' is 5.8, then a change in Xby one unit will bring out a change in yby 5.8 units. The positive value of 'b' indicates the increase in the value of y. It is associated with the increase of Xwhile a neaative value will tell the decrease in i/with an increase in X.

Procedure

1. Plot a graph between two variables taking independent variable on X-axis and dependent variable on Y-axis.

Find out the values of a and b using the equations given earlier. For drawing the line of best (regression line), find out any two values of y associated with corresponding value of xby using the equation:

2. Plot these two values on the graph on which all the points of original values have been put.
3. Make a straight line intersecting through these two points to get the fittest regression line.

Find out regression equation from the following data for 7 fishes of a species

Solution:
The following table is prepared first:

Equation (i) Zy = Sx.a + Sx.b or y = a+ bx
Equation (ii) £x/= Sx.a + 'L~.b orxy=x.a+b.~

Putting the values in the above formula:

18.3 =a+118.5/7

Zr=18.3

£X= 118.5

316.55 = 118.5a+2048.23b

SX/=316.55

SX= 118.5

SX2 = 2048.23

Multiply equation (i) with 118.5

=18.3x118.5= 118.5a+118.5xl-18.5b

2168.55 = 118.5a+14042.25b . ..(iii)

Subtracting equation (ii) from (iii) (2168.55-316.55) = (118.5-118.5) a+(14042.25 - 2048.23)b (2168.55-316.55) = 0 + (14042.25 - 2048.23)b

$$b = \frac{2168.55-316.55}{14042.25-2048.23} = \frac{1852}{11994.02}$$

b=0.16

Now put the value of b in equation (i)

18.3 =7a+0.16x118.5

18.3=7a+18.96

7a= 18.3 -18.96 = -0.66

a = -0. 6 = - 0.0943

Therefore, the Regression equation y=a+ bx

y=-0.0943+0.16x Ans.

Precautions while Using Regression Lines and Correlation Coefficient

1. Regression lines should not be extended beyond the range of data.

2. While predicting, the value of the dependent characteristic corresponding to the value of independent characteristic should lie between two observed values.

3. In interpreting a correlation coefficient, the two characteristics need not be directly related. They may be influenced by some other common factor.

4. The correlation coefficient gives the intensity of relationship only when the nature of relationship is linear. This is not possible for relationships other than linear.

5. lfx and y are independent, the coefficient of correlation is zero, but if r = 0, it is not necessary that x and y are independent. It means x and y may be merely uncorrelated.

6. Regression and correlation can be worked out only when data is reasonably homogeneous. In case the dots in the scatter diagram show clustering in two or more groups, the results obtained are spurious because of heterogeneity of data.

7. A spurious correlation may be observed even though x and y are independent of each other. It is found only in those cases where X and Y, though independent, vary roughly in the same way over a long period. Yule described such cases as the 'nonsense correlations'.

Direct method

(i) Regression Coefficient of X on Y (*bxy*) = $$\frac{\Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{n}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}$$

(ii) Regression Coefficient of Y on X (*byx*) = $$\frac{\Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

Find the regression coefficient (bxy) of X on Y when correlation coefficient is + 0.8. (S. D. of X series = 3.5 and S. D. of Y series = 2.5)

$$byx = r \times \frac{\text{SD of X series}}{\text{SD of Y series}}$$

$$= (0.8)\ \frac{3.5}{2.5} = \frac{2.8}{2.5} = 1.12 \quad \textbf{Ans.}$$

10.1.6 Differences between Regression Analysis and Correlation Analysis:

Regression analysis helps in proposing a mathematical model for ascertaining the relationship between two or more variables. Such models can be used to make predictions for one variable if the value of other variable is known. The ideas of regression were first elucidated by the English scientist. Sir Francis Galton, on reports of his research on heredity first in sweet pea and later in human features. Correlation analysis, on the other hand, is concerned with measuring the strength of the relationships between variables. When we compute measures of correlation from a set of data, we are interested in the degree of correlation between variables. Again, the concepts and terminology of correlation analysis originated with Galton.

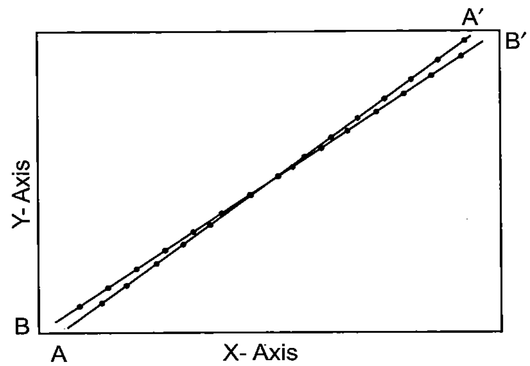


Fig.41 :Line of best fit.

In Fig.41, the scatter diagram shows two possible straight lines A-A'and B-B' that represent association between two variables in a given data. Regression analysis helps in determining which one of these two lines best represents the relationship. The line that best represents association between two variables in a given data is called line of best fit. In statistics, the equation of regression .line is usually represented as:

$$y = ax + b$$

Where, a is the slope,
b is the y-intercept

y is the 'y hat' that gives the predicted y value for a given x value

x is the x-intercept.

By the method of least squares analysis, the values of a and b are determined with accuracy and the equation of regression line y = ax + b best represents the relationship between the two variables. This regression line, therefore, is called the line of best fit.' The value of a and b can be calculated with the following equations:

When regression coefficient is y on x

$$a = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \text{ or } = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$b = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

EXERCISES

A. Numerical on Scatter Diagram

1. Draw a scatter diagram for the data given below and comment on the nature of correlation:

Series X	1	2	3	4	5
Series Y	20	16	13	11	10

2. From the heights of a group of fathers and sons given in the table below, construct a scatter diagram. Is the correlation positive or negative?

Heights of Fathers (inches)	65	63	67	64	68	62
Heights of Sons (inches)	68	66	68	65	69	66

3. Following are the heights and weights of 10 students of a class:

Heights (inches)	62	72	68	58	65	70
Weights (kg)	50	65	63	50	54	60

Draw a scatter diagram and indicate whether the correlation is positive or negative.

B. Numerical on Karl Pearson's Coefficient of Correlation

4. Find the coefficient of correlation between the variables-X and Y using Karl Pearson's method:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

5. Calculate the coefficient of correlation between X and Y series from the following data:

$N=15, X=25, \quad Y=18, \Sigma xy=122$

$\sigma_x = 3.01, \qquad \sigma_{xy} = 3.03$

6. From the following table, calculate the coefficient of correlation by Karl Pearson's method:

X	6	2	10	4	8
Y	9	11	7	8	7

Arithmetic Means of X and Y series are 6 and 8 respectively.

10.1.7 Statistical Packages:

SYSTAT:

SYSTAT is a statistics and statistical graphics software package, developed by Leland Wilkinson in the late 1970s, who was at the time an assistant professor of psychology at the University of Illinois at Chicago.

MINITAB :

Minitab is a statistics package developed at the Pennsylvania State University by researchers Barbara F. Ryan, Thomas A. Ryan, Jr., and Brian L. Joiner in 1972. It began as a light version of OMNITAB 80, a statistical analysis program by NIST. Statistical analysis software such as Minitab automates calculations and the creation of graphs, allowing the user to focus more on the analysis of data and the interpretation of results. It is compatible with other Minitab, LLC software.

SPSS:

SPSS Statistics is a software package used for interactive, or batched, statistical analysis. Long produced by SPSS Inc., it was acquired by IBM in 2009. The current versions (2015) are named IBM SPSS Statistics. The software name originally stood for Statistical Package for the Social Sciences (SPSS), reflecting the original market, although the software is now popular in other fields as well, including the health sciences and marketing.

S-PLUS:

S-PLUS is a commercial implementation of the S programming language sold by TIBCO Software Inc. It features object-oriented programming capabilities and advanced analytical algorithms. Due to the increasing popularity of the open source S successor R, TIBCO Software released the TIBCO Enterprise Runtime for R (TERR) as an alternative R interpreter.

NCSS:

NCSS is a statistics package produced and distributed by NCSS, LLC. Created in 1981 by Jerry L. Hintze, NCSS, LLC specializes in providing statistical analysis software to researchers, businesses, and academic institutions. It also produces PASS Sample Size Software which is used in scientific study planning and evaluation. The NCSS package includes over 250 documented statistical and plot procedures. NCSS imports and exports all major spreadsheet, database, and statistical file formats.

BMDP:

BMDP was a statistical package developed in 1965 by Wilfrid Dixon at the University of California, Los Angeles. The acronym stands for Bio-Medical Data Package, the word package was added by Dixon as the software consisted of a series of programs (subroutines) which performed different parametric and nonparametric statistical analyses. BMDP was originally distributed for free. It was later sold by Statsols, who originally was a subsidiary of BMDP, but through a management buy-out formed the now independent company Statistical Solutions Ltd, known as Statsols. BMDP is no longer available as of 2017. The company decided to only offer its other statistical product nQuery Sample Size Software.

GenStat:

Genstat (General Statistics) is a statistical software package with data analysis capabilities, particularly in the field of agriculture. Genstat is used in a number of research areas, including plant science, forestry, animal science, and medicine,[7] and is recognized by several world-class universities and enterprises.

STUDY QUESTIONS

Short Questions

1. **What is Chi-square test?**
2. **Mention the types of t-test**
3. **What is Regression?**
4. **What is SPSS?**

Long Questions

1. **Write about ANOVA.**
2. **Give an account on Correlation**
3. **Describe the Regression analysis**

BLOCK-IV : BIOINFORMATICS

Unit-XI :

Structure

11.1.Introduction:

Bioinformatics is an interdisciplinary subject involving molecular biology, genetics, computer science, statistics with data intensive and large-scale biological problems to be addressed from a computational point of view. The most common problems are modeling biological processes at the molecular level and making inferences from collected data. A bioinformatics solution usually involves the following steps 1) Collect statistics from biological data, 2) Build a computational model. 3)Solve a computational modeling problem & 4) Test and evaluate a computational algorithm. This chapter gives a brief introduction to bioinformatics by first providing an introduction to biological terminology and then discussing some classical bioinformatics problems organized by the types of data sources. The Identification of homologs, multiple sequence alignment, searching sequence patterns, and evolutionary analyses can be studied by using Sequence analysis of DNA and Protein. . Protein structures are three-dimensional data and the associated problems are structure prediction (secondary and tertiary), analysis of protein structures for studying the function, and structural alignment. Gene expression data is usually represented as matrices and analysis of microarray data mostly involves statistics analysis, classification, and clustering approaches. Biological networks such as gene regulatory networks, metabolic pathways, and protein-protein interaction networks are usually modeled as graphs and graph theoretic approaches are used to solve associated problems such as construction and analysis of large-scale networks.

11.2. Medical Informatics :

Medical informatics is found at the intersection of healthcare and technology. It is where skills in both medical and computer sciences come together in an effort to improve healthcare and patient outcomes. Professionals in this hybrid field draw on expertise from

both disciplines to put technology to its best use in patient care, clinical and research settings. This field deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine.

11.2.1 Types of Work in Medical Informatics:

Medical informatics professionals are tasked with using information technology to its greatest advantage in the healthcare industry. This means they are responsible for such tasks as Creating, maintaining or facilitating new ways for medical facilities and practices to keep electronic health records (EHR), Improving communication between healthcare providers and facilities to ensure the best patient outcomes, Storing, managing and analyzing data for research, This field also includes assisting with complex, technology-dependent research, such as that involved in human genome sequencing.

11.2.2 Specialties in the field of Medical Informatics:

Medical and computer science professionals who wish to transition into this field with a master's or doctoral degree and students heading straight into this arena will find a number of specialties exist. They include

a) **Medical Bioinformatics:** Practitioners in this specialty are concerned with storing, retrieving, sharing and helping analyze biomedical information for research and/or patient care. Subspecialties include chemical, nursing and dental informatics.

b) **Public health informatics:** This specialty involves the use of technology to guide how the public learns about health and health care while also ensuring access to the latest medical research. Professionals also ensure public health practices have access to the information they need.

c) **Organizational informatics:** The focus here is ensuring a smooth flow of communication within a healthcare organization.

d) **Social informatics:** These specialists study the social aspects of computer science while gaining insights into how information technology affects social environments and how social environments affect information technology.

e) **Clinical informatics:** This is the application of informatics and information technology for clinical research and patient care. Professionals leverage information technology for medical education, patient education and students, among others.

11.3. Cheminformatics :

Cheminformatics (also known as **chemoinformatics**, (**or) chemical informatics**) is the use of computer and informational techniques applied to a range of problems in the field of chemistry. These *in silico* techniques are used in, for example, pharmaceutical companies in the process of drug discovery. These methods can also be used in chemical and allied industries

in various other forms. Chemo informatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization.

11.3.1 Basics of Cheminformatics :

Cheminformatics combines the scientific working fields of chemistry, computer science and information science for example in the areas of topology, chemical graph theory, information retrieval and data mining in the chemical space. Cheminformatics can also be applied to data analysis for various industries like paper and pulp, dyes and such allied industries.

11.4. Pharmacoinformatics:

Pharmacoinformatics is new emerging information technologies like neuroinformatics, immunoinformatics, Metabolomics, chemo-informatics, toxico-informatics, cancer informatics, genome informatics, proteome informatics, biomedical informatics are basic tools provided for the purpose of drug discovery. Drug discovery and development requires the integration of multiple scientific and technological disciplines. These include chemistry, biology, pharmacology, pharmaceutical technology and extensive use of information technology. The latter is increasingly recognised as Pharmacoinformatics. The main idea behind the field is to integrate different informatics branches (e.g. bioinformatics, chemoinformatics, immunoinformatics, etc.) into a single platform, resulting in a seamless process of drug discovery. The first reference of the term "Pharmacoinformatics" can be found in the year of 1993. The first dedicated department for Pharmacoinformatics was established at the National Institute Of Pharmaceutical Education And Research, S.A.S. Nagar, India in 2003.[2] This has been followed by different universities worldwide including a program by European universities named the European Pharmacoinformatics Initiative (EuroPIN).

Pharmacy informatics can be thought of as a sub-domain of the larger professional discipline of health informatics. Health informatics is the study of interactions between people, their work processes and engineered systems within health care with a focus on pharmaceutical care and improved patient safety. For example, the Health Information Management Systems Society (HIMSS) defines pharmacy informatics as, "the scientific field that focuses on medication-related data and knowledge within the continuum of healthcare systems - including its acquisition, storage, analysis, use and dissemination - in the delivery of optimal medication-related patient care and health outcomes"

UNIT-XII

Structure

12.1 Current researches in bioinformatics

12.2 Application of Bioinformatics in cancer detection

12.3 Drug Targets

12.1.Current Researches:

Most cellular tasks are performed not by individual proteins, but by groups of functionally associated proteins, often referred to as modules. In a protein assocation network modules appear as groups of densely interconnected nodes, also called communities or clusters. These modules often overlap with each other and form a network of their own, in which nodes (links) represent the modules (overlaps). We introduce CFinder, a fast program locating and visualizing overlapping, densely interconnected groups of nodes in undirected graphs, and allowing the user to easily navigate between the original graph and the web of these groups. We show that in gene (protein) association networks CFinder can be used to predict the function(s) of a single protein and to discover novel modules. CFinder is also very efficient locating the cliques large sparse graphs.

12.2.Application of Bioinformatics in Cancer Detections:

Cancer is a disease determined by several genetic and epigeneiic alter— ations. Due to technological advances in the omics discipl ines, cancer research is going through a revolution. The technological advances that lead to the post- genome era have allowed molecular biologists to make meticulous studies on the DNA (genome), the rilRNA (transcriptomc) and the protein sequences (prolcomc). [f11t161iVCS that intend to describe cancer in a global dimension are providing an opportunity fni investigators to have more useful d2tH to analyze arifi integrate ie novel ways. Despite the practical difficulties, a gruming number of projects are being developed with the aim to integrate information about samples, protocols, and data front multiple sources. Cancer bioinformatics dears with the organization and analysis of lhc data so that important trends and pattcrns can be identified the ultimate goal being the discovery of new therapeutic and/or diagnostic protocols for caocei. In this chapter, we will discuss snrne aspects of this ievoluiion giving a special emphasis on Bioinformatics. Further more, we will discuss how the omics data is being analyzed and used to transform the way cancer patients an treated. One of the biggest challenges facing cancer researchers is that the disease varies so much from person to person. Even the same type of cancer – lung, brain, breast, colon, and so on – can be subtly different. This means that a therapy that works well in one patient may have no effect in another. So researchers in the UK brought in the big guns – ***bioinformatics***. Cancer Research UK has set up seven British centers to start collecting 9,000 tumor samples from a wide range of cancer patients to create a DNA database. Researchers will extract DNA from these tumors and scan them for a series of key genes involved in tumor development. The results will then be cross-checked against a range of cancer treatments, to create a map of which treatments work best for

cancers associated with which particular genes. This is based on the concept of pharmacogenomics: certain genes predispose people to respond to certain drugs in certain ways. We can already test a cancer patient for a single gene, knowing how tumors with that gene respond to a particular drug. However currently we don't have a way of testing a broad panel of genes. And to compound the problem, we don't have a way of quickly and accurately sharing information between labs in the same city, across the country or internationally. Again, enter the power of bioinformatics. With the proposed cancer DNA database, a doctor might analyze a patient's tumor sample and prescribe a tailored treatment plan within a very short period of time, perhaps as little as two weeks. As Professor Matthew Seymour, director of the National Cancer Research Network (NCRN) in the UK, recently stated, "We have to get clever about how to target drugs. Medications for cancer have to be personalized because no two cancers are identical." Bioinformatics research is increasing at an exponential rate. DNA sequences are available to anyone with an Internet connection – along with free bioinformatics tools to explore sequence data, predict the presence of genes, and compare features shared between organisms. The DNALC has been working in DNA sequencing and bioinformatics for over a decade, developing intuitive, visually appealing computer tools for teachers and students to quickly learn the rudiments of gene analysis and integrate bioinformatics with biochemistry labs.

12.3.Drug Targets :

Target identification and validation are among the most important steps In developing a new drug. A target can cover a range of biological entities, including proteins genes and RNA. A good target must be druggable, i.e. accessible to the drug and upon binding, must elicit a biological response, that is measurable both in vitro and vivo. Additionally a good target must be efficacious and meet clinical and commercial needs. In a standard Drug Target Discovery project, we combine patient data (typically case-control data) with in-depth data mining and understanding of protein pathways and networks to investigate and identify novel drug targets. The text mining includes our proprietary inBio Know™ data mining platform and for protein-protein interaction networks we apply our proprietary PPI network database, inBio Map™

Project Planning:

We work closely with our clients in the planning of each project. Each project is tailored specifically to generate the most valuable outcome, taking available data, in-house capabilities, timelines etc into account.

Duration:

The duration of Drug Target Identification projects is usually from 4 to 12 months, depending on the complexity of the project.

Close project cooperation and communication:

A key factor in the successful outcomes with our partners, is the close interaction and communication we have from planning the projects and all the way through project execution and presentation of the project deliverables. We typically have bi-weekly Webex update/planning meetings with the client and have face-to-face meetings every 2-3 months during the project. At the end of the project, we present the project deliverables and conclusions

in a face-to-face meeting. All face-to-face meetings take place at our clients' facility, unless another location is desired.

Deliverables:

In our comprehensive project report, we deliver a prioritized list of potential targets, with evidence-based documentation for each target. Apart from the drug target list, the report also documents the project steps and details key project findings. Additionally, we present potential next steps in the project for discussion, if relevant. In the scoping and preparation phase for all our client projects, we work closely with our clients to ensure the optimal data foundation, since this is a central parameter for successful project outcomes. For drug target projects this often includes: • Conducting a thorough data survey of existing data • Evaluating data quality/experimental background • Including biological signal from relevant omics data • Checking for consistency among data sets and known disease mechanism. When project scoping and collection of all relevant data have been finalized, we move into the data analysis and interpretation phase. In a typical drug target identification project a process involves: • Applying Intomics standard QC procedures to evaluate data quality • Combining data types to enhance the biological signal and applying our proprietary network biology data to: › Discover novel biological relationships › Establish prioritized biological hypothesis for identified drug targets based on consistency across data • Establishing drug targets from network-guided machine learning/AI. Upon project conclusion, we ensure that the project deliverables have actionable outputs. Historically we have provided: • Prioritized list of targets • Hypothesis linking drug targets to disease mechanisms

The deliverables, that Intomics provides at the end of a project or project phase, are of very high quality and are well documented. In the project scoping and throughout the execution of the projects, we work closely with our clients to ensure that the results from the projects are as actionable for the individual clients as possible. In drug target identification projects our deliverables often include: • A ranked list of identified drug targets with associated, supporting evidence • A link of the prioritized drug targets and the associated biological context/hypothesis • An extensive project report (data quality, methods, summary of key findings, list of additional interesting findings, description of role of targets, target properties and druggability assesment) • And, when relevant, well-documented motivation for additional project work that would increase the value of the project

STUDY QUESTIONS:

Short Questions

1. **Write briefly about the current researches in bio informatics**
2. **What are Drug targets**

Long Questions

1. **Write about the applications of bioinformatics in cancer detection**

UNIT-XIII :

Structure

13.1.Animal genome diversity:

Domestication of animals was an essential step in human demographic and cultural development. Together with the domestication of plant species it laid the foundation of agriculture as we know it today. During the subsequent history of livestock, the main evolutionary forces of mutation, selective breeding, adaptation, isolation and genetic drift have created an enormous diversity of local populations. In the last centuries, this has culminated in the formation of many well-defined breeds used for a variety of purposes with differing levels of performance. During the last decades, development of and increased focus on more efficient selection programmes have accelerated genetic improvement in a number of breeds. Artificial insemination and embryo transfer have facilitated the dissemination of genetic material. In addition, progress in feed technology has allowed optimal nutrition, while enhanced transport and communication systems have led to uniform and strictly controlled production environments. As a result, highly productive breeds have replaced local ones across the world. This development has led to growing concerns about the erosion of genetic resources (Food and Agriculture Organization of the United Nations. As the genetic diversity of low-production breeds is likely to contribute to current or future traits of interest they are considered essential for maintaining future breeding options. According to the FAO, 20% of the roughly 7600 breeds, belonging to 18 mammalian species and 16 avian species, reported worldwide are at risk, and 62 breeds became extinct within the first 6 years of this century. Effective management of farm animal genetic resources (FAnGR) requires comprehensive knowledge of the breeds‘ characteristics, including data on population size and structure, geographical distribution, the production environment, and within- and between-breed genetic diversity. Integration of these different types of data will result in the most complete representation possible of biological diversity within and among breeds, and will thus facilitate effective management of FAnGR.

13.2. Introduction to DNA and Protein sequences analysis:

When we design a recombinant DNA construct, it is important to know the potential restriction endonuclease recognition sites both in the vector and the insert. Finding these sites is a relatively simple bioinformatics task using online programs such as NEBCutter or Restriction Mapper. It is also useful to draw a map of the recombinant, online programs such as pDRAW, Bio Edit or Snap Gene. Another typical bioinformatics task is the design of oligonucleotide primers for sequencing, polymerase chain reaction (PCR) or site-specific *in*

*vitro*mutagenesis. This can also be achieved by online programs such as Primer3, Oligo or OligoCalc.

13.2.1 DNA Sequences:

One has to search a similarity to learn if our sequenced DNA can be found in a public nucleotide database (i.e. it has already been cloned by others) and/or whether it is evolutionally related (i.e. homologous) to other sequences. In a simple similarity search, one can compare a sequence with sequences found in an entire nucleotide database (see later the BLAST program), while for a homology search the method of choice is multiple sequence alignment by the ClustalW program. By comparing either nucleotide or amino acid sequences we can find homologs. If these are from different species (that had a common ancestor) but have identical or similar functions they are called **orthologs**; while those homologs that are found in the same organism and originate from a gene duplication event followed by divergent evolution within the species are called **paralogs**. We will not cover the construction of evolutionary trees in this e-book—one can learn about these in bioinformatics or evolutionary biology courses.

If we sequence a DNA clone, the first bioinformatics analysis is a similarity search against a nucleotide database. The most widely used similarity search program accessible on the internet is **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool), which will be described here and will be used by the students during the laboratory practice. The BLAST program is available online at several servers including the one at NCBI:

BLAST uses a heuristic algorithm that makes it possible to search a huge database in a very short period of time by using a **query sequence**. The high speed of the algorithm stems from the fact that the query sequence is divided into short „words" that are used, instead of the full-length sequence, during the alignment process. These words are searched in the database first (called „seeding", i.e. finding the best local alignments). The most relevant hits are then scored with the help of a scoring matrix, extended to neighboring words, and finally assembled and compiled into a final list of similarity hits. It is important that the query sequences must be in the so-called **FASTA format** (FASTA was a previously popular but much slower similarity search program). The FASTA format is shown in Figure 41.

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
        (the first line can be omitted)
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

Fig 41 :The FASTA sequence format

If we want to search using a nucleotide query sequence within a nucleotide database, we can use the **BLASTN** version of the program. If we have an amino acid sequence, we can search a protein database by the **BLASTP** version of the program. The **BLASTX** version of the program

translates a nucleotide sequence in all six reading frames (three on each strand) and allows searching a protein database. Finally, with the **TBLAST** subprogram, we can search against a translated nucleotide database using either a protein (TBLASTN) or a nucleotide (TBLASTX) query sequence. These similarity search options are summarisedin Table 6.

Table 6: Search possibilities in the BLAST program

Program	Query Sequence	Target database
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide in 6 reading frame	Protein
TBLASTN	Protein	Nucleotide in 6 reading frame
TBLASTX	Nucleotide in 6 reading frame	Nucleotide in 6 reading frame

The result of a BLAST analysis is a list a sequences from the searched database that show significant similarity to the query sequence. Besides the sequence identifiers of the similar sequence hits in the database, the final list of alignments contains a score number and a statistical significance number, the ***E*-value**. The *E*-value is a parameter that describes the number of hits one can expect to see by chance when searching a database of a particular size. It decreases exponentially as the score (S) of the match increases. Essentially, the *E*-value describes the random background noise. The lower the *E*-value, or the closer it is to zero, the more "significant" the match ($E > 0.01$ is usually considered to reflect a homologous, i.e. evolutionarily-related sequence). The score value is calculated based on the alignment, taking into account the gaps and the similarity of the amino acids at the aligned positions. The most often used similarity matrix (an amino acid substitution matrix) is the **BLOSUM** (**Blo**cks **Su**bstitution **M**atrix) matrix. The numbers within a BLOSUM are “log-odds” scores that measure, in an alignment, the logarithm of the ratio of the likelihood of two amino acids appearing with a biological sense and the likelihood of the same amino acids appearing by chance.

13.2.2. Protein Sequences:

The wide range of *in silico* analysis possibilities of protein sequences is summarised in Figure 42. Note that many of these analyses can be performed also with nucleic acid sequences. Sequences can be compared to each other and to full databases. The physical and structural/functional properties of polypeptide chains can be predicted via this analysis. Sequence comparisons (alignments) were described in the previous section (BLAST and ClustalWprograms). During the so-called profile analysis, the analysed sequences are compared to secondary databases that contain information about protein structural families, structural and functional domains, modules, phosphorylation, glycosylation and other posttranslational modification consensus sequences. Many online programs are available on the internet that can search secondary databases. For instance, the InterProScan profile analysis program can be used to search the InterPro secondary database (in fact it is a „superdatabase” of several individual derived databases) maintained by the EBI. Another example is

the PhosSitePlus database that can be searched by any query sequence to predict phosphorylation or other posttranslational modification sites.

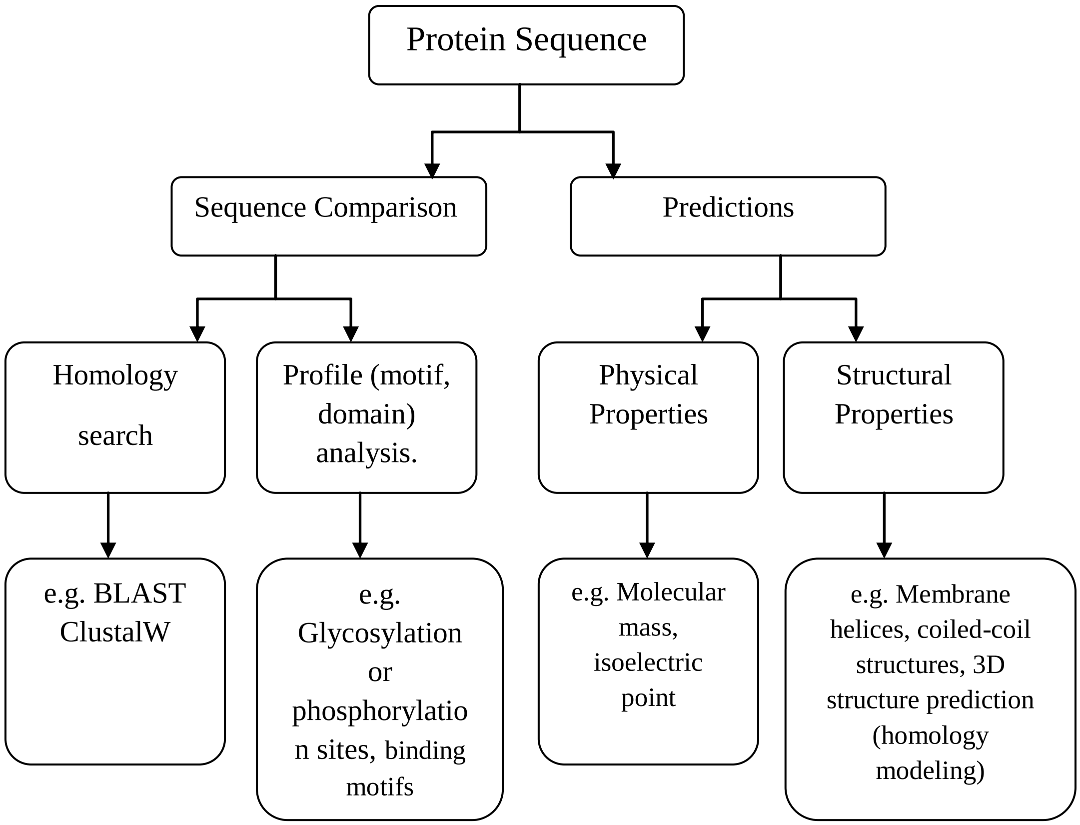


Figure.42 : The wide range of *in silico* analysis possibilities of protein sequences. (Most of these options are also available for nucleic acid sequences.)

STUDY QUESTIONS:

Short Questions

1. **What is an Animal genome diversity**
2. **Define BLASTA**

Long Questions

1. **Give an account on DNA Sequence Analysis**
2. **Explain Protein Sequence Analysis**
3. **Discuss the concepts of biological; data base**

UNIT-IV

Structure

14.1 Phylogenetic Analysis:

14.1.1 PHYLIP:

PHYLIP is a comprehensive phylogenetic analysis package created by Joseph Felsenstein at the University of Washington. This package can do many of the phylogenetic analyses available in the literature today. Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites, distance matrices, and 0/1 (binary) discrete characters. Installation PHYLIP is freely available from. The editing of the Clustal alignment format is easier than the editing of PHYLIP-format, and Clustal will readily read in the .aln format, if you later need to convert the edited sequences into some other format. Any multiple sequence alignment can also be manually reformatted with a text editor. The format requirements for PHYLIP are rather stringent, and any deviation will result in a program that hangs, usually with the error message Unable to allocate memory. The file must confirm to the following (Felsenstein, PHYLIP documentation): 1. The file begins with the information about the number of sequences and the number of nucleotides or amino acids in the alignment. 2. The sequence names must be exactly 10 characters long. Spaces can be added to the end of shorter names to make them this length. Do not use Tab characters for this. 3. Gaps must be indicated by - . 4. Missing data or missing information (no sequence) is indicated by ?. 5. Spaces between the alignment blocks are allowed. This normally makes the alignment more readable. Spaces are usually inserted into the alignment every 10 bases or amino acids. 6. Blanks will be ignored, and so will numerical digits. This allows GENBANK and EMBL sequence entries to be read with minimum editing. Example of the formatted sequences: 5 100 Rabbit ?????????? ?????????C CAATCTACAC ACGGG-GTAG GGATTACATA Human AGCCACACCC TAGGGTTGGC CAATCTACTC CCAGGAGCAG GGAGGGCAGG Opossum AGCCACACCC CAACCTTAGC CAATAGACAT CCAGAAGCCC AAAAGGCAAG Chicken GCCCGGGGAA GAGGAGGGGC CCGGCGG-AG GCGATAAAAG TGGGGACACA Frog GGATGGAGAA TTAGAGCACT TGTTCTTTTT GCAGAAGCTC AGAATAAACG TTTGGATGGT AG---GATAT GGGCCTACCA TGGCGTTAAC GGGT-AACGY TTTCGACGGT AA---GGTAT TGGCTTACCG TGGCAATGAC AGGT-GACGG TTTCGACGGT AA---GGTAT TGGCTTACCG TGGCAATGAC AGGT-GACGY TTTCGACGGT AA---GGTAT TGGCTTACCG TGGCAATGAC AGGT-GACGG TTTCGATGGT AA---GGTAT TGGCTTACCG TGGCAATGAC AGGT-GACGG Possible ambiguities (such as N, Y or R nucleotides) are also handled correctly, and do not cause trouble. 8 Font files In order to be able to use the tree-drawing tools, the font files need to be in the same folder as the Drawtree or Drawgram

program(s). If you are using PHYLIP on a PC from the same folder it was installed in, you should not encounter any troubles. However, this is not strictly necessary, just remember to copy the font files with tree drawing programs to the same folder. Or, better still, copy and rename your favorite fontfile as fontfile and keep only it with the tree drawing programs. There are six different fonts available: font1 simple sans-serif Roman font2 medium quality sans-serif Roman font3 high quality sans-serif Roman font4 medium quality sans-serif Italic font5 high quality sans-serif Italic font6 Russian Cyrillic Running PHYLIP programs The programs are used in a sequential way. The output from the first program is used as an input in the next program. The trick is to know how to use the programs in suitable combinations. See the flow charts in the end of this book for some suggestions. In Windows, the PHYLIP programs can be invoked by double-clicking on the icon or by typing the name of the program on the command line. It is advisable to use programs from the command line, because then you will be better able to see, e.g., the error messages that might appear. In NT-line Windows versions (NT, 2000 and XP) the DOS prompt, i.e., command line, can be invoked from Start -> All Programs -> Accessories -> Command Prompt. Most PHYLIP programs run in the same way. The input for a program is taken from a file called infile - if the program does not find this file it then asks the user to type in the file name of the data file. The results are written in a file called outfile. Some programs may write both outfile and a file called outtree or plotfile. Because most of the programs use the default names for the input and output files, you need to be sure to rename the files you want to save before proceeding to further analysis. Otherwise you risk losing your results. For example, you get a distance matrix (outfile) from the program Dnadist, but you want to try different settings for the matrix calculations. Then, before doing the matrix calculation again, rename outfile to Dnadist_out_F84 or something similar, so that you can tell different analysis results apart after you have ceased to work. 9 Essential programs Here is a list of the programs that can be used for the molecular sequence data analysis. The programs are divided into the method categories. The choice of the correct analysis method is left for the user. Distance methods These programs are intended to be used sequentially. First a distance matrix is calculated by Dnadist or Protdist program from the multiple sequence alignment. The matrix is then transformed into a tree by Fitch, Kitsch or Neighbor program. Programs Dnadist and Protdist create a file outfile. Before running Fitch, Kitsch ot Neighbor, outfile should be renamed, either as infile ot with another file name. Fitch, Kitsch and Neighbor programs create both outfile and outtree. Dnadist DNA distance matrix calculation Protdist Protein distance matrix calculation Fitch Fitch-Margoliash tree drawing method without molecular clock Kitsch Fitch-Margoliash tree drawing method with molecular clock Neighbor Neighbor-Joining and UPGMA tree drawing method Character based methods These programs read in the sequence alignment, and produce either one or multiple trees in the output files outfile and outtree. Dnapars DNA parsimony Dnapenny DNA parsimony using branch-and-bound Dnaml DNA maximum likelihood without molecular clock Dnamlk DNA maximum likelihood with molecular clock Protpars Protein parsimony Proml Protein maximum likelihood Resampling tool This program reads in a sequence alignment, and generates a specified number of random samples into a file outfile. These random samples are usually used in subsequent analysis as a sequence alignment file with the option M ("use multiple datasets") turned on. Seqboot Generates random samples by bootstrapping or jack-knifing Tree drawing These programs draw a tree from the specifications in the Newick-format. For example, the specification can be in a file produced by the program Dnaml. The Newick file outtree 10 produced by Dnaml should be renamed to intree before visualizing the tree. Drawgram and Drawtree produce a file plotfile, whereas Retree saves the result in a file outtree. Drawgram Draws a rooted tree Drawtree Draws an unrooted tree Retree Interactive tree-rearrangement Consensus trees This program constructs a consensus tree from multiple trees. For example, Dnapars can produce multiple trees, which can be summarized by the program

Consense. Also the results of bootstrapping are summarized by the program Consense as a majority rule tree. Consense Draws consensus trees from multiple trees Tree distances This program computes, e.g., a topology-based distance between two or more trees. The distance can be used to assess or compare the results from different analyses. Treedist Computes distances between trees based on tree topology 11 Quick start. Here a DNA sequence data is used as an example.

14.1.2 Clustal W :

Clustal performs a global multiple sequence alignment by a stepwise process. In step 1 it performs pair wise alignment of all the sequences provided by the user. In step 2 the scores obtained for the pair wise alignment are used to produce a phylogenetic tree and in step 3 the phylogenetic tree used as a guide to align sequences sequentially. Thus the most closely related sequences are aligned first, then additional sequences are added one by one to a profile of an existing MSA. The scoring of gaps is done in the manner different from the followed for a pair wise alignment. ClustalW calculates gaps in a Novel way. The graphical version of ClustalW provides a versatile environment for doing MSA of sequences. Alignments can also be produced in a profile mode. Profile mode is typically used when MSA is already known for a set of sequences and when MSA is already known for a set of sequences and when one wants to align a sequence of one MSA to another MSA. This is very useful feature for finding conserved domains. ClustalX has graphical user interface. Clustal was developed by Higgins and Sharp in 1988 and many improved versions were developed later. ClustalW is the latest version of clustal with W standing for Weighing to represent the ability of the program to provide weights to the sequence and program parameters.

Patterns in multiple sequence alignments Multiple sequence alignment is the process of aligning several related sequences, showing the conserved and unconserved residues across all of the sequences simultaneously. These conserved and unconserved residues form a pattern that can often be used to retrieve sequences that are distantly related to the original group of sequences. These distant relatives are extremely helpful in understanding the role that the groups of sequences play in the process of life.

Global multiple sequence alignments Global multiple sequence alignments ore sequence alignments that require the participations of all sequence residues. A multiple sequence alignments shows the residue juxtaposition across the entire set of sequences, thus showing the conserved and unconserved residues across all of the sequences simultaneously. Progressive par wise approach The progressive paiwise approach relies on exhaustive pairwise alignments between all of the sequences to produce a measure of sequence relatedness. From this measure an algorithm (UPGMA in Pileup, Neighbor Joining in ClustalW) is used to develop a joining order. This joining order corresponds to a tree that is used to proceed with the multiple sequence alignment. It should be noted that this tree is not an evolutionary tree. The tree is shown in Fig 43.

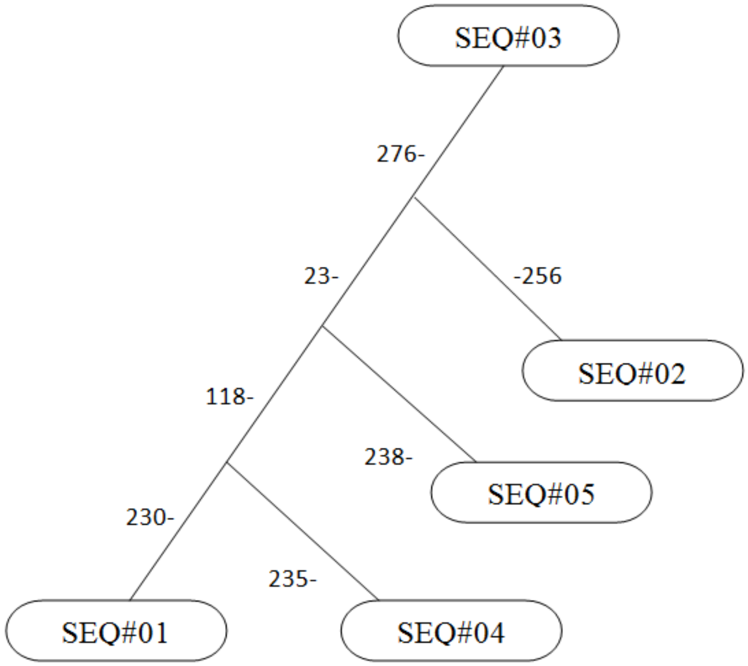


Fig.43 : Tree of multiple sequence alignment. (Progressive pair wise approach)

Five Sequences have been taken for tree formation seq01,04,have close relationship that seq 05 and 02 but a far relationship to seq 03. After the joining order has been determined, sequences close tp each other are aligned first. In the example above, Seq#01 and SEQ#04 are the first two sequences to be aligned. The third sequence, SEQ#05, is then aligned with the two previously aligned sequences, SEQ#01 and SEQ#02 is then aligned, followed by SEQ#03. While this approach produces adequate results for many sets of sequences, the alignment produced by the procedure will vary depending on the joining order:

[SEQ#03 + SEQ#02] + [SEQ#05]+[SEQ#04] + [SEQ#01]

may not produce the same alignment as joining the sequences in the original; order.

[SEQ#01 + SEQ#04] + [SEQ#05] + [SEQ02] + [SEQ#03]

The advantages top this approach are that it require only modest computer resources and that it is capable of aligning hundreds of sequences.

STUDY QUIESTIONS:

Short Questions

1. **Define PHYLIP**

2. **What is ClustalW**

Long Questions

1. **Explain about Phylogenetic Analysis in Bio informatics**

MODEL QUESTION PAPER

35043- BIOPHYSICS, BIOSTATISTICS AND BIOINFORMATICS

Time: 3 hours Max marks: 75

Part - A (10 X 2 = 20)

Answer all questions

1. What is the structure of an atom.
2. Define Redox Potential
3. What is Natural Radiation?
4. What is an Isotope?
5. Define Biostatistics
6. Define Random Sampling
7. What is a Polygon?
8. What is a Range?
9. What is Pharmacoinformatics
10. Define BLASTA

Part - B (5x5 = 25 Marks)

11.a. Describe the various chemical bonds.

(Or)

b. Write about role of NADP and NAPH in Biological Systems.

12.a. Discuss the Energy States of Atoms.

(or)

b. Give an account on Autoradiography.

13.a. Give an account on collection of data**.**

(or)

b. Explain the concept of Sampling.

14. a. Write about the various types of Mean.

(or)

b. Enumerate the characteristics features of a normal distribution?

15. a. Explain Protein Sequence Analysis

(or)

b. Describe briefly about ClustalW

Part – C (3 x 10 = 30 Marks)

16. Give an account on polymerization of organic molecules

17. Write about the properties of Light

18. Explain the three measures of central tendency

19. Describe probability and hypothesis testing

20. Explain the DNA Sequence analysis

REFERENCE BOOKS

1. Nolting, B (2006) Methods in modern biophysics, Springer, Berlin
2. Agarwal, S. K. (2005) Advanced biophysics, APH Publishing Corporations, India
3. Daniel, M. (2004) Basic biophysics, Agrobios publications, India
4. Goutham, N, Pattabi, S. (2001). Biophysics, Narossa Publishing company, New Delhi.
5. Daniel, W. W. (2007) Biostatistics, Wiley publishers, USA
6. Zar (2006) Biostatistical analysis, Dorling Kindersley Pvt. Ltd , India.
7. Bailey, N.T.J. (1997), Statistical Methods in Biology, III Ed., Cam. University Press, N.Y.
8. Neil A. Weiss (1995). Introductory statistics. Addison Wesley Publishing company, Inc.
9. McCleery, R.H and Watt, T.A.,(2007). Introduction to statistics for biology.3rd Ed., Chapman & Hall / CRC Press.
10. Mount, D.,(2004). "Bioinformatics: Sequence and Genome Analysis"; Cold Spring Harbor Laboratory Press, New York.
11. Lesk, A.M., (2002), "Introduction to Bioinformatics", First Edition, Oxford University Press, UK.
12. Lukas, K., Buehler, Hooman, H.Rashidi, (2000)"Bioinformatics Basics: Application in Biological Science and Medicine"; CRC Press.
13. Dubrin.R, S.Eddy, A.Korgh and G.Hitchison, (2003), Biological Science Analysis, Cambridge University Press, Eighth edition.